

# 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**

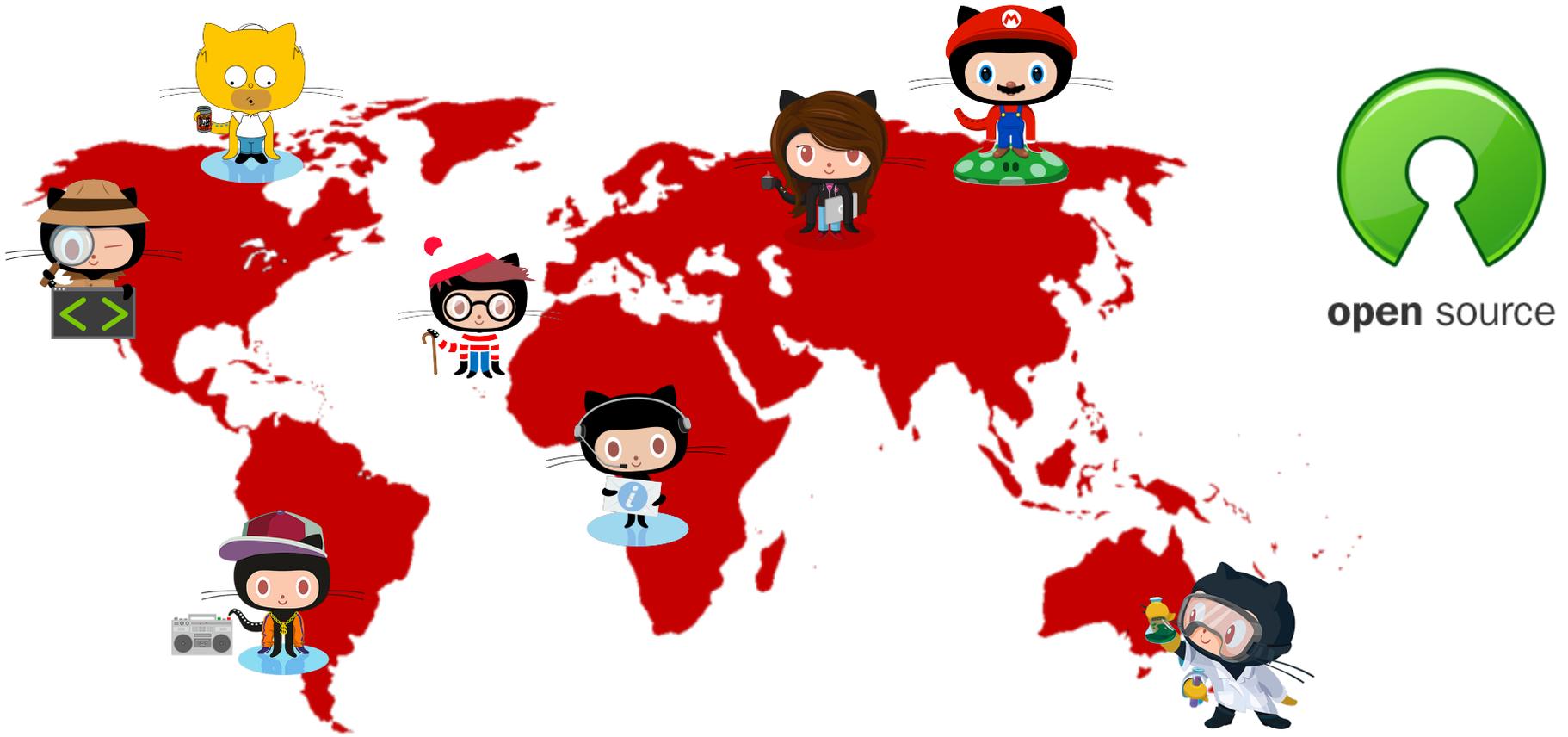
# Collaboration is Everywhere



# Globally Distributed Software Development



# 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



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



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



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



Problem

Intervention

Evaluation



Help software developers  
to better collaborate



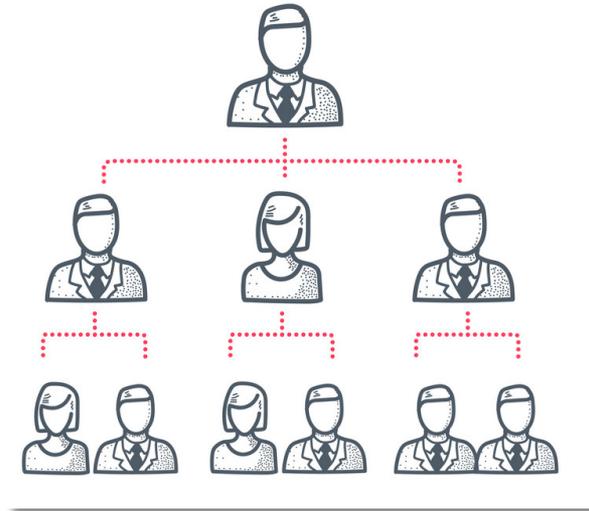
- + Advances in tooling & SE principles
- + Insights from other disciplines
- + Mix a wide range of research methods



# Traditional Collaboration Model



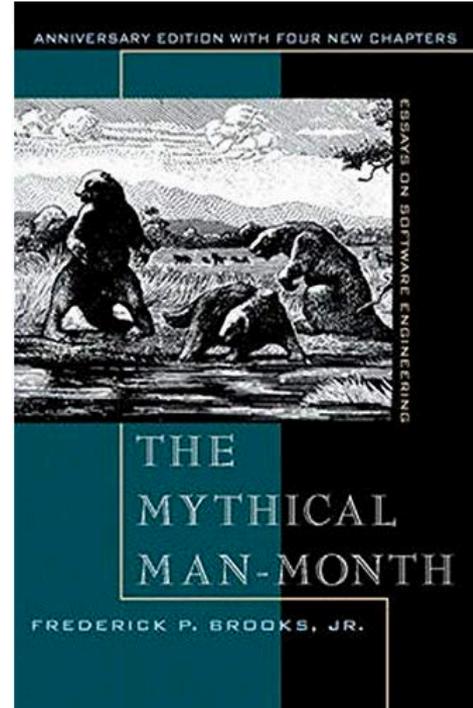
# Traditional Collaboration Model



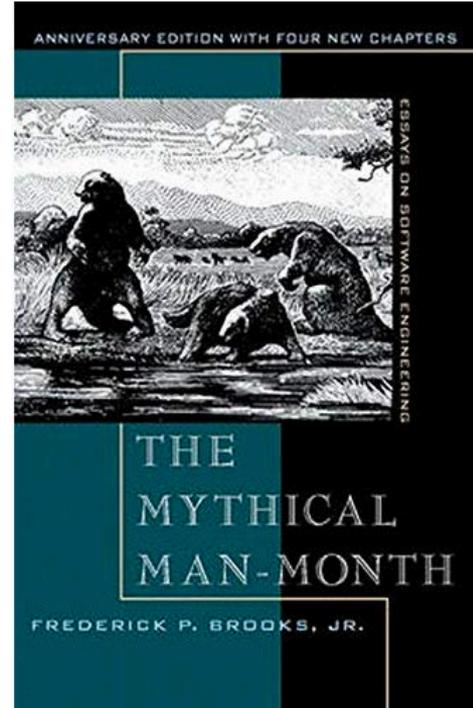
# Traditional Collaboration Model



# Traditional Collaboration Model



# Traditional Collaboration Model



# CSCW

Computer Supported  
Cooperative Work

# Fork-Based Development Changed Everything

# Fork-based Dev. Changed Everything



**GitHub**



**Bitbucket**



**GitLab**

# Traditional Collaboration Model

90's



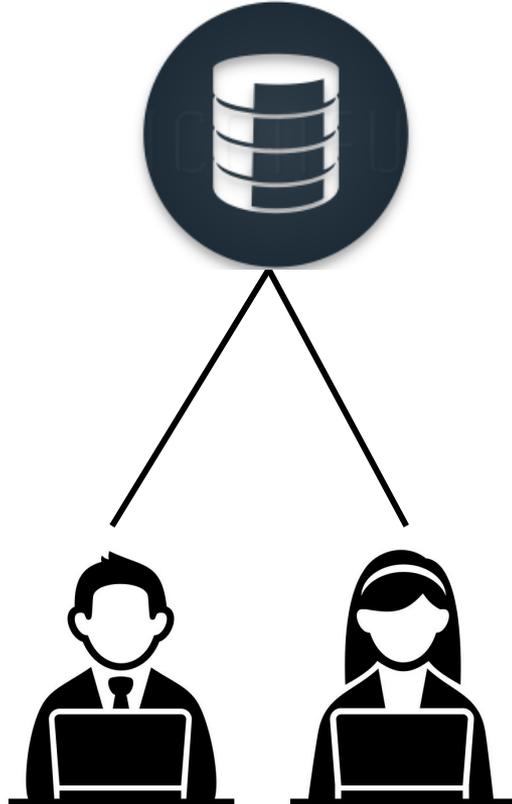
open source

# Traditional Collaboration Model

90's



open source

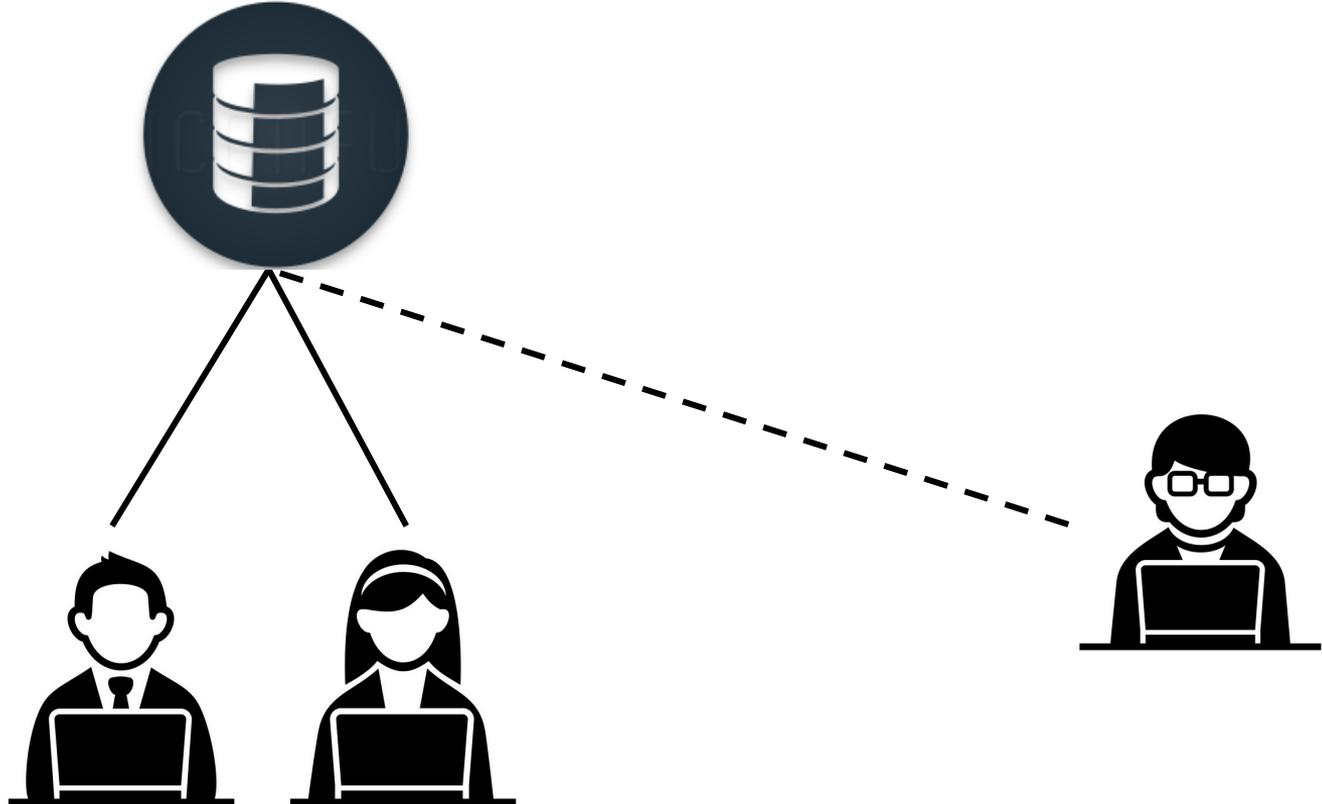


# Traditional Collaboration Model

90's



open source

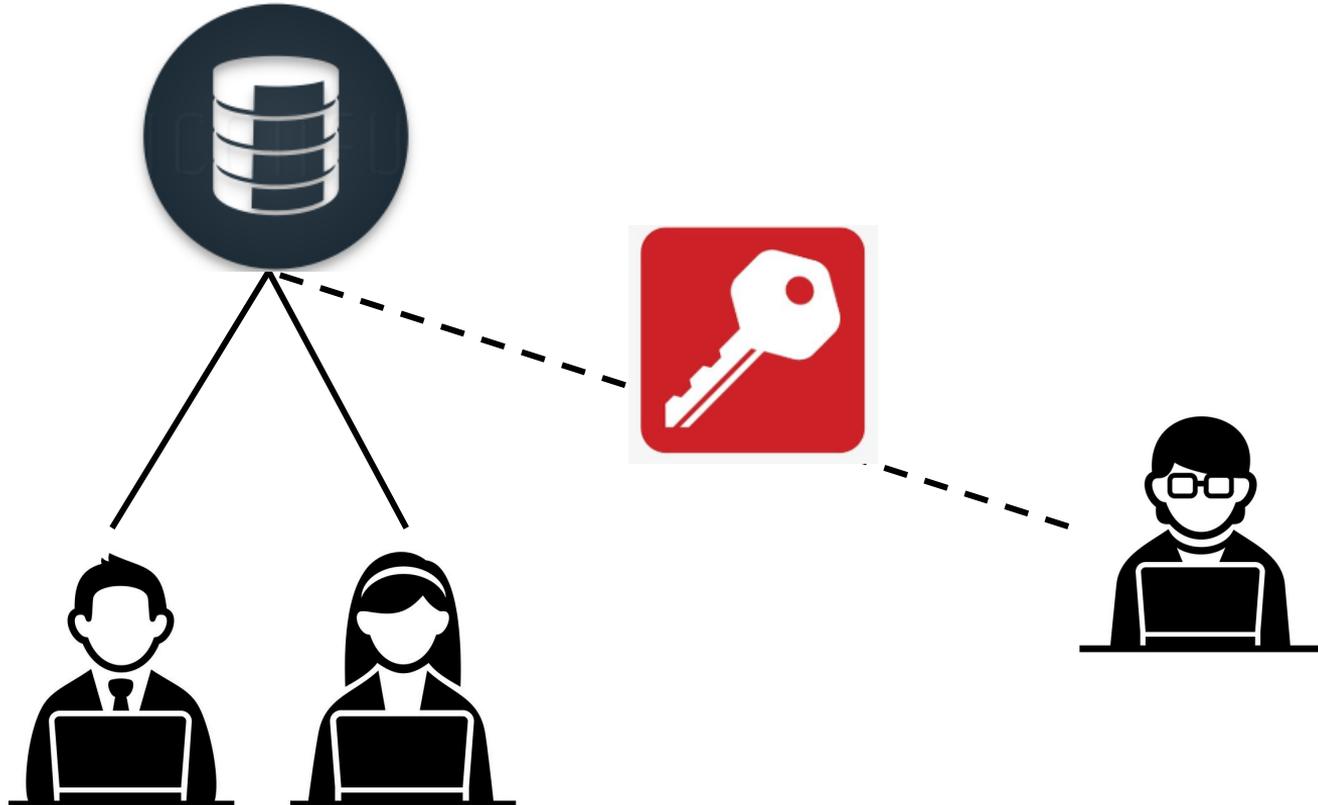


# Traditional Collaboration Model

90's



open source



# Traditional Collaboration Model

90's



open source



# 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 OMAP 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 is:

```
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 -rN linux-2.4.20.orig/Documentation/Configure.help linux-2.4.20/Documentation/Configure.help
--- linux-2.4.20.orig/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
```

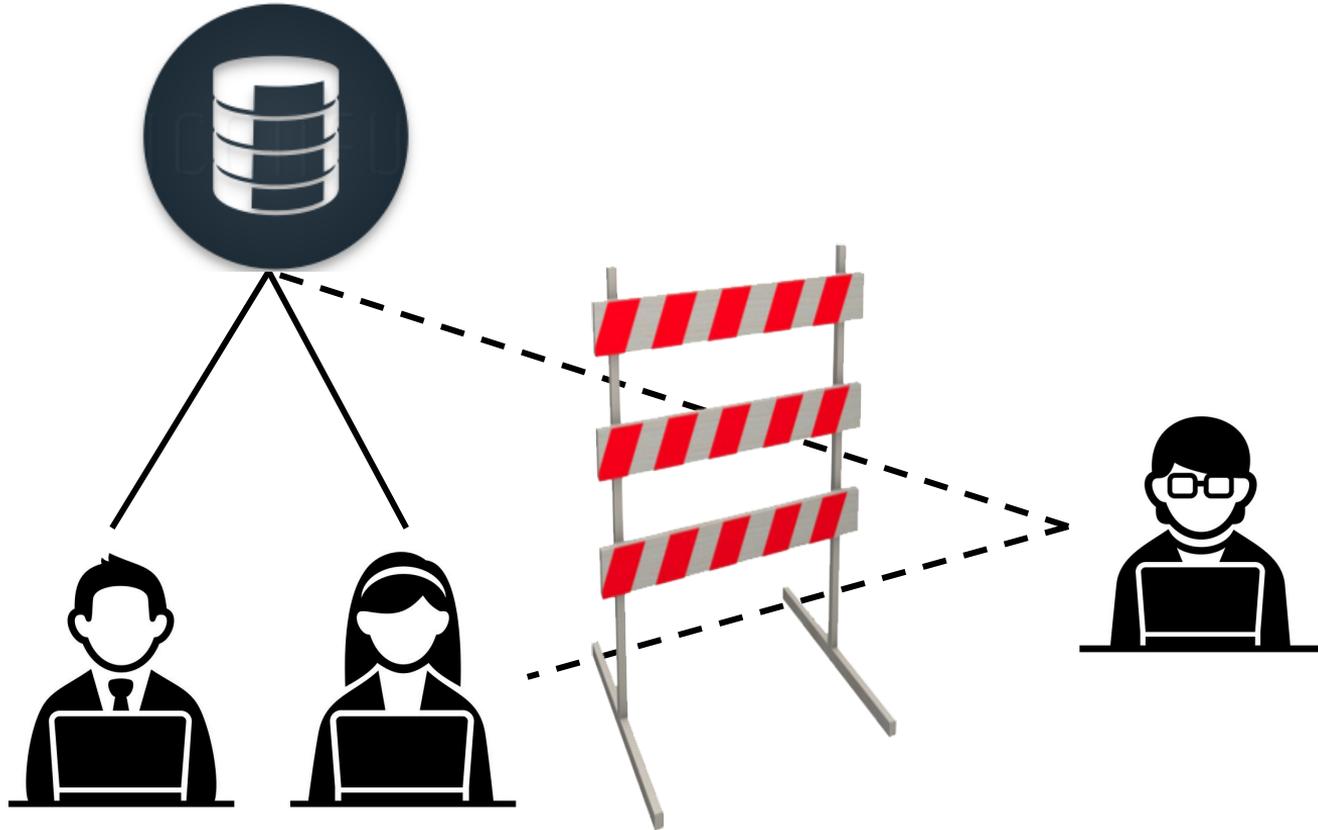


# Traditional Collaboration Model

90's



open source



# Fork-based Development



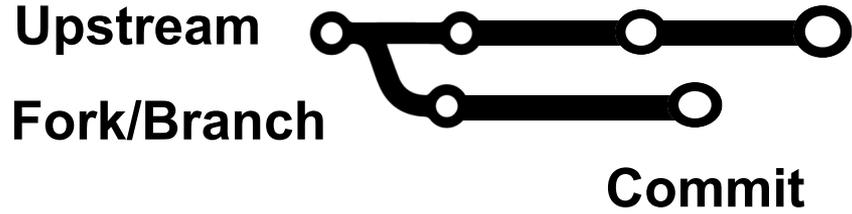
# Fork-based Development



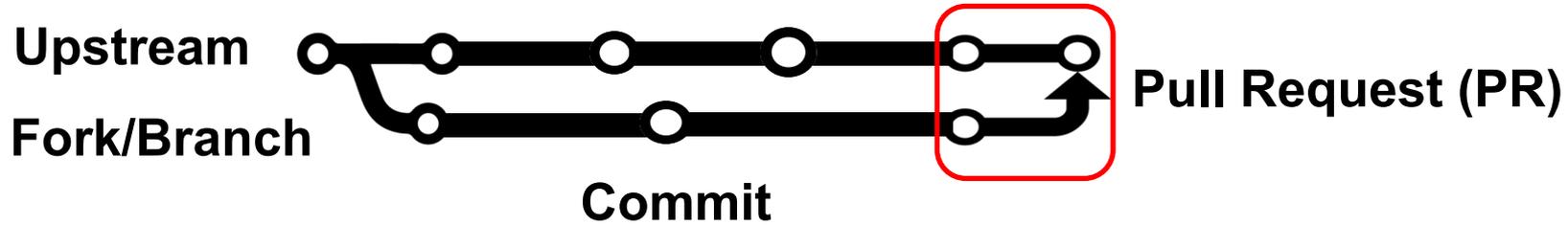
# Fork-based Development



# Fork-based Development



# Fork-based Development



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

**Pull Request / Merge Request**

# Fork-based Dev. Lowers Entry Barriers

The screenshot shows the GitHub repository page for scikit-learn. At the top, the repository name is 'scikit-learn / scikit-learn'. To the right, there are statistics for 'Used by' (86.4k), 'Watch' (2.3k), 'Star' (39.1k), and 'Fork' (19.2k). The 'Fork' button is circled in red. Below the repository name, there are links for 'Code', 'Issues' (1,398), 'Pull requests' (722), 'Actions', 'Projects' (17), 'Wiki', 'Security', and 'Insights'. The repository description is 'scikit-learn: machine learning in Python' with a link to 'https://scikit-learn.org'. There are tags for 'machine-learning', 'python', 'statistics', 'data-science', and 'data-analysis'. Below the description, there are statistics for '25,081 commits', '20 branches', '0 packages', '106 releases', '1,571 contributors', and a 'View license' link. At the bottom, there are buttons for 'Branch: master', 'New pull request', 'Create new file', 'Upload files', 'Find file', and 'Clone or download'. A recent commit is shown with the message '5 authors DOC clarifications on the release process (#15759)' and a link to the commit. Below the commit, there are three folders: '.binder', '.circleci', and '.github', each with a corresponding commit message and date.

scikit-learn / scikit-learn

Used by 86.4k Watch 2.3k Star 39.1k Fork 19.2k

<> Code Issues 1,398 Pull requests 722 Actions Projects 17 Wiki Security Insights

scikit-learn: machine learning in Python <https://scikit-learn.org>

machine-learning python statistics data-science data-analysis

25,081 commits 20 branches 0 packages 106 releases 1,571 contributors View license

Branch: master New pull request Create new file Upload files Find file Clone or download

5 authors DOC clarifications on the release process (#15759) Latest commit 1382831 6 minutes ago

.binder	MAINT: simpler binder requirements.txt (#14832)	5 months ago
.circleci	[MRG] MNT Updates pypy to use 7.2.0 (#15954)	last month
.github	MNT remove tag help wanted in doc template (#16122)	11 days ago

# Fork-based Dev. Lowers Entry Barriers

The screenshot displays the GitHub interface for the `scikit-learn/scikit-learn` repository. At the top, the main repository statistics are shown: 86.4k users, 2.3k watches, 39.1k stars, and 19.2k forks. The `Fork` button is circled in red. Below this, a smaller repository view for `shuiblue/scikit-learn` is shown, also with a circled `Fork` button. The main repository page shows the `machine-learning` branch selected, with 25,081 commits and 20 branches. A table of recent commits is visible, including one by 5 authors regarding DOC clarifications on the release process.

scikit-learn / scikit-learn

Used by 86.4k Watch 2.3k Star 39.1k Fork 19.2k

Code

shuiblue / scikit-learn  
forked from scikit-learn/scikit-learn

Watch 0 Star 0 Fork 19.2k

Code Pull requests 0 Actions Projects 0 Wiki Security Insights Settings

scikit-learn: machine learning in Python <https://scikit-learn.org> Edit

Manage topics

25,081 commits 20 branches 0 packages 106 releases 1,571 contributors View license

Branch: master New pull request Create new file Upload files Find file Clone or download

This branch is even with scikit-learn:master. Pull request Compare

5 authors DOC clarifications on the release process (scikit-learn#15759) Latest commit 1382831 9 minutes ago

.binder	MAINT: simpler binder requirements.txt (scikit-learn#14832)	5 months ago
.circleci	[MRG] MNT Updates pypy to use 7.2.0 (scikit-learn#15954)	last month
.github	MNT remove tag help wanted in doc template (scikit-learn#16122)	11 days ago

# Fork-based Dev. Lowers Entry Barriers

Upstream  
Fork/Branch



Pull Request (PR)

scikit-learn / scikit-learn

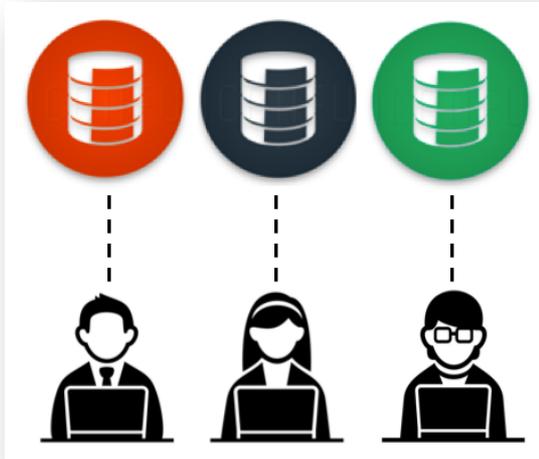
Used by 86.4k Watch 2.3k Star 39.1k Fork 19.2k

Code Issues 1,39 Pull requests 723 Actions Projects 17 Wiki Security Insights

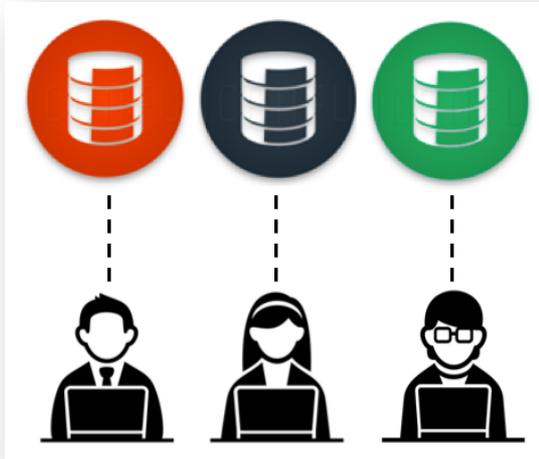
Filters is:pr is:open Labels 29 Milestones 4 New pull request

723 Open	8,461 Closed	Author	Label	Projects	Milestones	Reviews	Assignee	Sort
[MRG] Fix FutureWarning in plot_partial_dependence_visualization_api	✓	#16256	opened 2 minutes ago by kssling					
[MRG] Adding explained variances to sparse pca	✓	#16255	opened 1 hour ago by Batalex					
"Improved error message when plotting a not fitted tree."	✗	#16253	opened 1 hour ago by Rick-Mackenbach					2
ENH Add 'if_binary' option to drop argument of OneHotEncoder	✓	#16245	opened 23 hours ago by rushabh-v • Changes requested					24

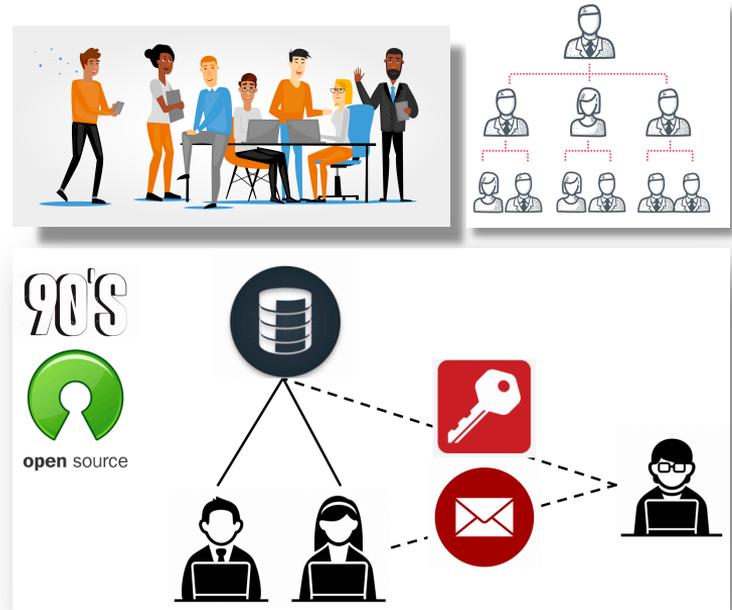
# Fork-based Development



# Fork-based Development



VS



# Fork-based Dev. Becomes Popular

#Forks	#GitHub Projects
>50	114,120
>500	9164
>1,000	2236
>5,000	198
>10,000	72
>100,000	2

[GHTorrent 2019-06]



# Fork-based Dev. Becomes Popular

#Forks	#GitHub Projects
>50	114,120
>500	9164
>1,000	2236
>5,000	198
>10,000	72
>100,000	2

[GHTorrent 2019-06]



**open** source

# Fork-based Dev. Becomes Popular

#Forks	#GitHub Projects
>50	114,120
>500	9164
<b>&gt;1,000</b>	<b>2236</b>
>5,000	198
>10,000	72
>100,000	2

[GHTorrent 2019-06]



**open** source

# Fork-based Dev. Becomes Popular

**NETFLIX**



American  
Airlines



**GROUPON**



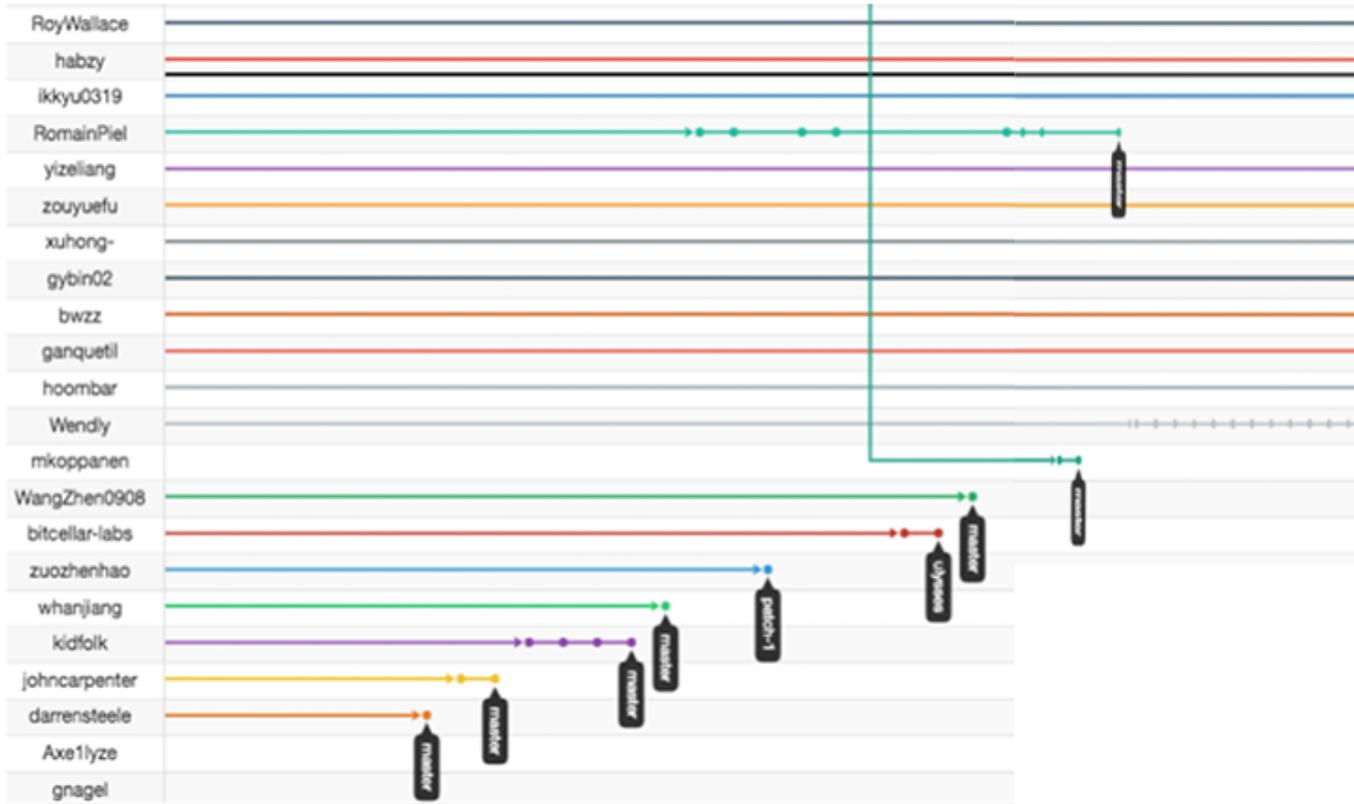
**Linked in**



**Companies**

But

# Problem -- Lost Contributions



# Problem -- Redundant Development



foosel commented on Aug 22, 2017

Owner



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



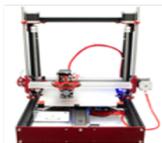
Noiredd commented on Nov 3, 2017

Member

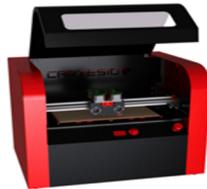
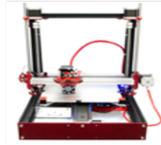


Duplicate of [#5869](#) and [#5972](#), partially also [#5879](#).

# Problem -- Fragmented Community



# Problem -- Fragmented Community



Behind the Scenes Bytes

## 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

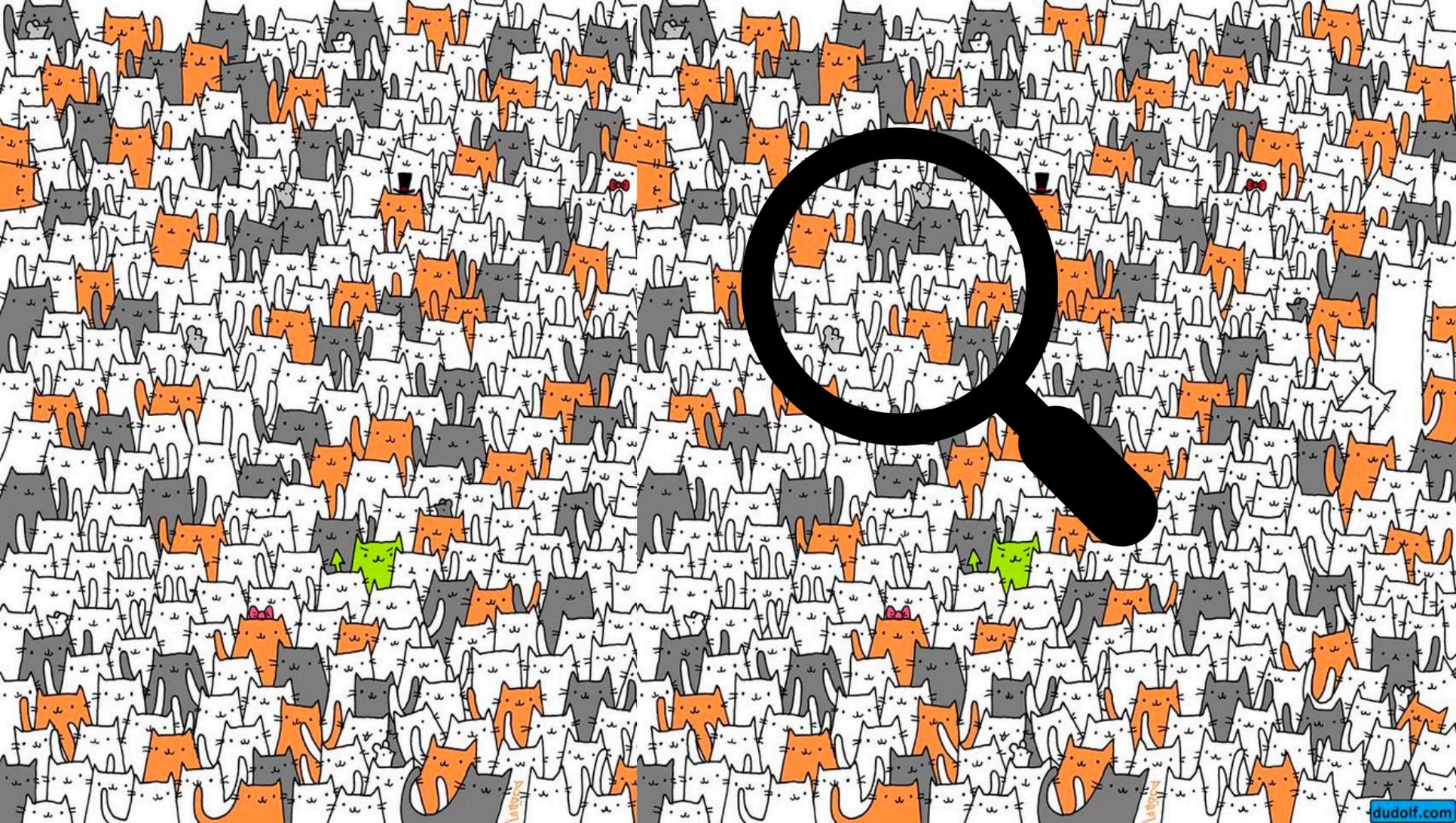
**Is duplicate always bad?**

# Similar Problems Happen in Industry

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].



- [1] Thorsten Berger, Divya Nair, Ralf Rublack, Joanne M Atlee, Krzysztof Czarnecki, and Andrzej Wąsowski. 2014. Three Cases of Feature-based Variability Modeling in Industry. In Proc. Int'l Conf. Model Driven Engineering Languages and Systems (MoDELS)
- [2] Anh Nguyen Duc, Audris Mockus, Randy Hackbarth, and John Palframan. 2014. Forking and Coordination in Multi-platform Development: A Case Study. In Proc. Int'l Symp. Empirical Software Engineering and Measurement (ESEM). ACM



# Problem

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

**Closed** Opened 4 years ago by  **Adriano Vieira**

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

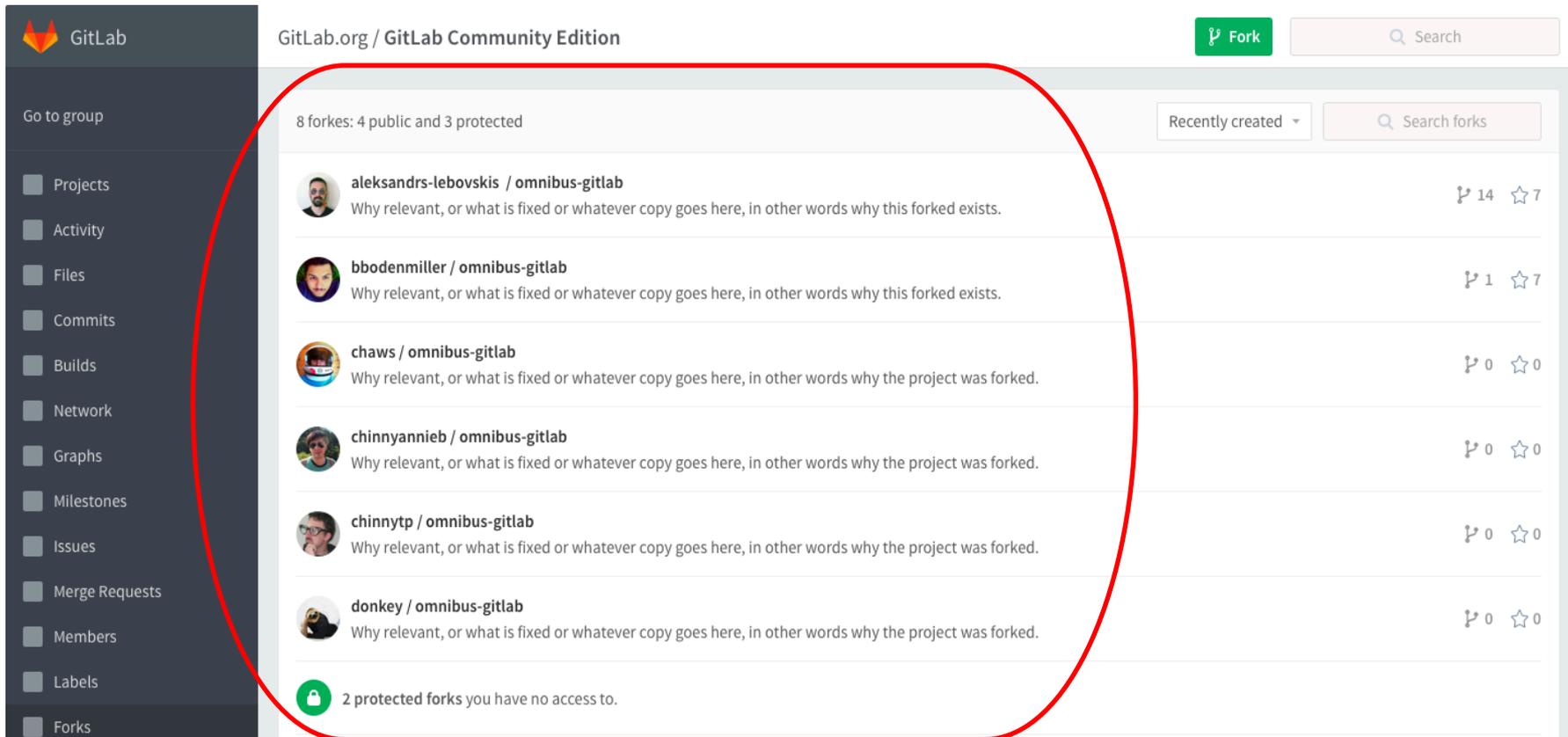
You have on the project home page a button which show us a quantity of forks from one project.

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

How could we see all forked projects of any project?

# Problem

## List of Forks



The screenshot shows the GitLab interface for the 'omnibus-gitlab' project. A red circle highlights the main content area, which includes a list of forks and a sidebar with navigation options. The sidebar on the left contains a 'Go to group' section and a list of navigation items: Projects, Activity, Files, Commits, Builds, Network, Graphs, Milestones, Issues, Merge Requests, Members, Labels, and Forks. The main content area shows the project name 'omnibus-gitlab' and a list of 8 forks (4 public, 3 protected). Each fork entry includes the forker's profile picture, name, and a description. The forks are: alexsandr-lebovskis (14 forks, 7 stars), bbodenmiller (1 fork, 7 stars), chaws (0 forks, 0 stars), chinnyannieb (0 forks, 0 stars), chinnytp (0 forks, 0 stars), and donkey (0 forks, 0 stars). At the bottom, it indicates 2 protected forks that the user has no access to. The top right of the page features a 'Fork' button and a search bar. The bottom right corner of the page shows the number '50'.

GitLab

GitLab.org / GitLab Community Edition

Fork

Search

8 forks: 4 public and 3 protected

Recently created

Search forks

 **alexsandr-lebovskis / omnibus-gitlab**  
Why relevant, or what is fixed or whatever copy goes here, in other words why this forked exists. 14 forks 7 stars

 **bbodenmiller / omnibus-gitlab**  
Why relevant, or what is fixed or whatever copy goes here, in other words why this forked exists. 1 fork 7 stars

 **chaws / omnibus-gitlab**  
Why relevant, or what is fixed or whatever copy goes here, in other words why the project was forked. 0 forks 0 stars

 **chinnyannieb / omnibus-gitlab**  
Why relevant, or what is fixed or whatever copy goes here, in other words why the project was forked. 0 forks 0 stars

 **chinnytp / omnibus-gitlab**  
Why relevant, or what is fixed or whatever copy goes here, in other words why the project was forked. 0 forks 0 stars

 **donkey / omnibus-gitlab**  
Why relevant, or what is fixed or whatever copy goes here, in other words why the project was forked. 0 forks 0 stars

 2 protected forks you have no access to.

50

# Problem

## Network View

Smoothieware / Smoothieware

Watch 196

Star 661

Fork 648

Code

Issues 7

Pull requests 12

Projects 0

Wiki

Insights

Pulse

Contributors

Community

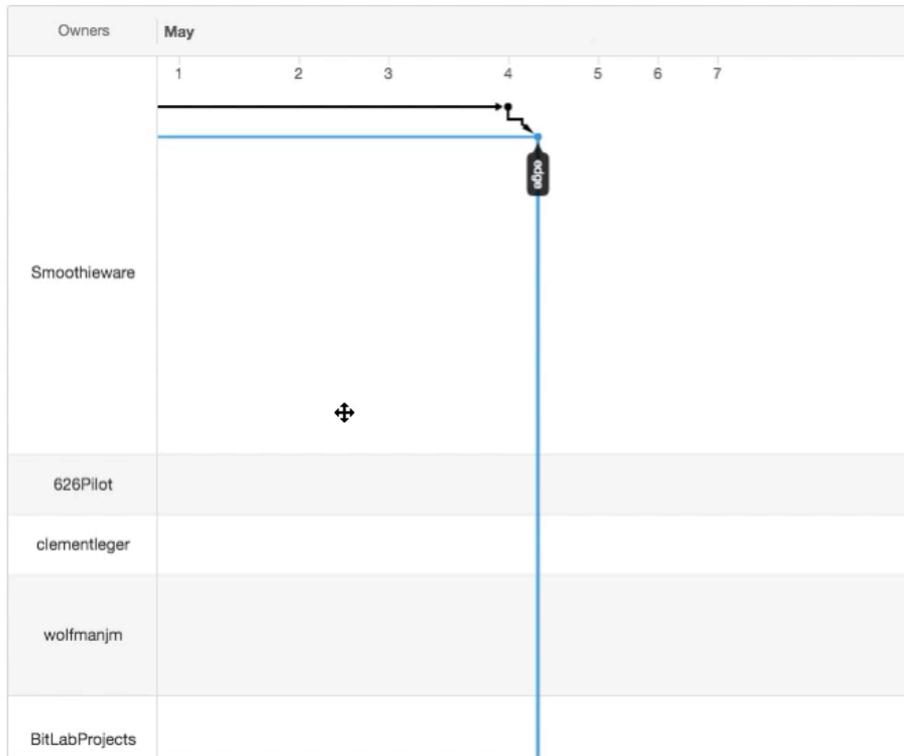
Commits

Code frequency

Dependency graph

**Network**

Forks



# Problem

## Network View

Smoothieware / Smoothieware

Watch 196

Star 661

Fork 648

Code

Issues 7

Pull requests 12

Projects 0

Wiki

Insights

Pulse

Contributors

Community

Owners May

5 6 7

Lack of  
Overview

berger

wolfmanjm

BitLabProjects

# 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



**Software  
Dev.**

**Distributed**

**Problem**

**Fork-Based**

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

# Improving Collaboration Efficiency



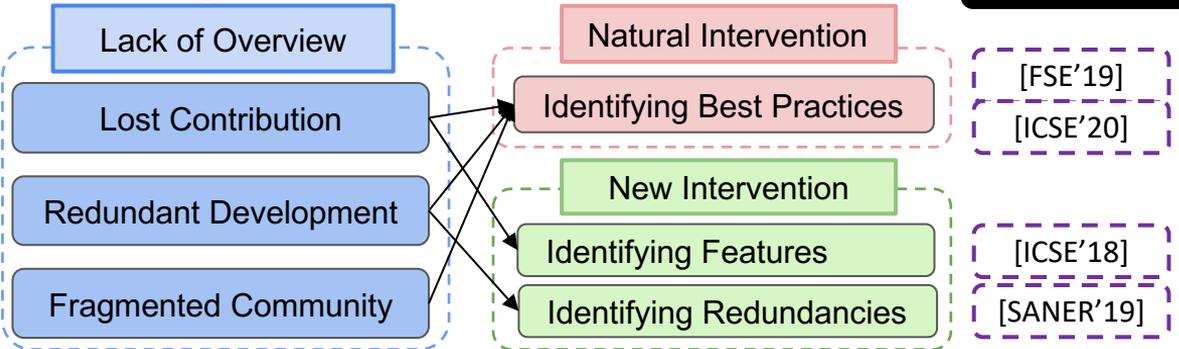
**Software  
Dev.**

**Distributed**

**Problem**

**Solution**

**Fork-Based**



# Improving Collaboration Efficiency



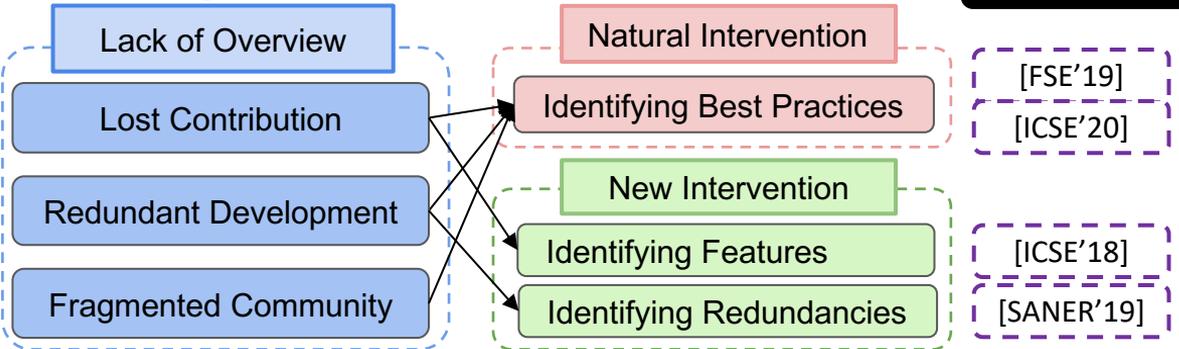
**Software Dev.**

**Distributed**

**Problem**

**Solution**

**Fork-Based**



# Problem

# Solution



Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Natural Intervention

Identifying Best Practices

[FSE'19]

[ICSE'20]

New Intervention

Identifying Features

[ICSE'18]

Identifying Redundancies

[SANER'19]

# Problem

# Solution



Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Natural Intervention

Identifying Best Practices

[FSE'19]

[ICSE'20]

New Intervention

Identifying Feature

[ICSE'18]

Identifying Redundancies

[SANER'19]

# Problem

# Solution



Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Natural Intervention

Identifying Best Practices

New Intervention

Identifying Features

Identifying Redundancies

[FSE'19]

[ICSE'20]

[ICSE'18]

[SANER'19]

# Research Statement

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

# Solution



Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Natural Intervention

Identifying Best Practices

New Intervention

Identifying Features

Identifying Redundancies

[FSE'19]

[ICSE'20]

[ICSE'18]

[SANER'19]

# Problem

# Solution



Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Natural Intervention

Identifying Best Practices

[FSE'19]

[ICSE'20]

New Intervention

Identifying Feature

[ICSE'18]

Identifying Redundancies

[SANER'19]

# Solution 1 – Identifying Natural Interventions

[FSE 2019]

## What the Fork: A Study of Inefficient and Efficient Forking Practices in Social Coding

Shurui Zhou

Carnegie Mellon University, USA

Bogdan Vasilescu

Carnegie Mellon University, USA

Christian Kästner

Carnegie Mellon University, USA

# Solution 1 – Identifying Natural Interventions

[FSE 2019]

## What the Fork: A Study of Inefficient and Efficient Forking Practices in Social Coding

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



VS



- Project proposal
- Resolve issues on the issue tracker

- Open for any contribution

# Projects are different



VS



- **Centralized Mgmt**
- **Upfront Coordination through Issue Tracker**

- **De-centralized Mgmt**
- **No Upfront Coordination**

# Coordination Mechanism Affects Forking Practices

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]

Organizational Theory



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



# Research Question

**What characteristics and practices of a project associate with efficient forking practices?**

# Research Method

Interviewing Stakeholders

Literature/Theory Search



Deriving  
Hypotheses

# Derive 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

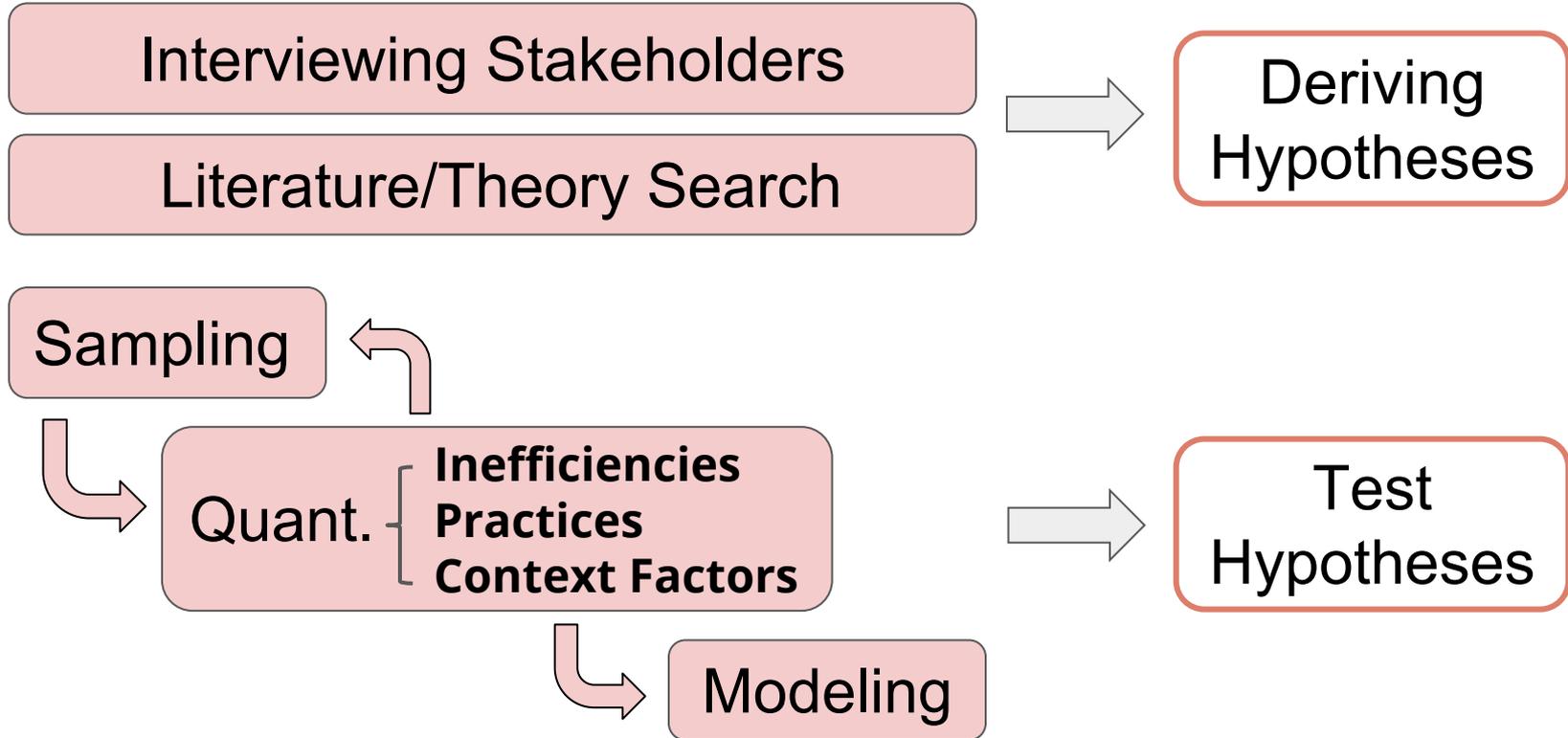


- + Advances in tooling & SE principles
- + Insights from other disciplines
- + Mix a wide range of research methods



Developmental Psychology

# Research Method



# Test Hypotheses

Sampling

Group	#fork	#projects	#projects in sample set
A	[3,000 , +]	231	200
B	[1,000 , 3,000)	847	300
C	[20 , 1,000)	116,532	1300

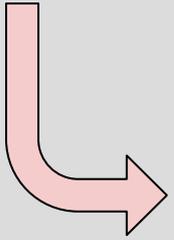
Quantifying { Inefficiencies  
Practices  
Context Factors

Multiple Regression Modeling

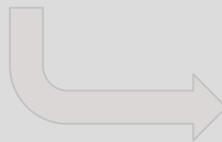
# Test Hypotheses

Sampling

Group	#fork	#projects	#projects in sample set
A	[3,000 , +]	231	200
B	[1,000 , 3,000)	847	300
C	[20 , 1,000)	116,532	1300



Quantifying {  
Inefficiencies  
Practices  
Context Factors

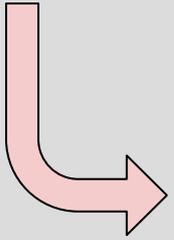


Multiple  
Regression  
Modeling

# Test Hypotheses

Sampling

Group	#fork	#projects	#projects in sample set
A	[3,000 , +]	231	200
B	[1,000 , 3,000)	847	300
C	[20 , 1,000)	116,532	1300



Quantifying {  
Inefficiencies  
Practices  
Context Factors

Hypo: Centralized Management → Larger portion of contributing forks

Regression 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

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

 Conversation 9

 Commits 7

 Checks 11

 Files changed 2



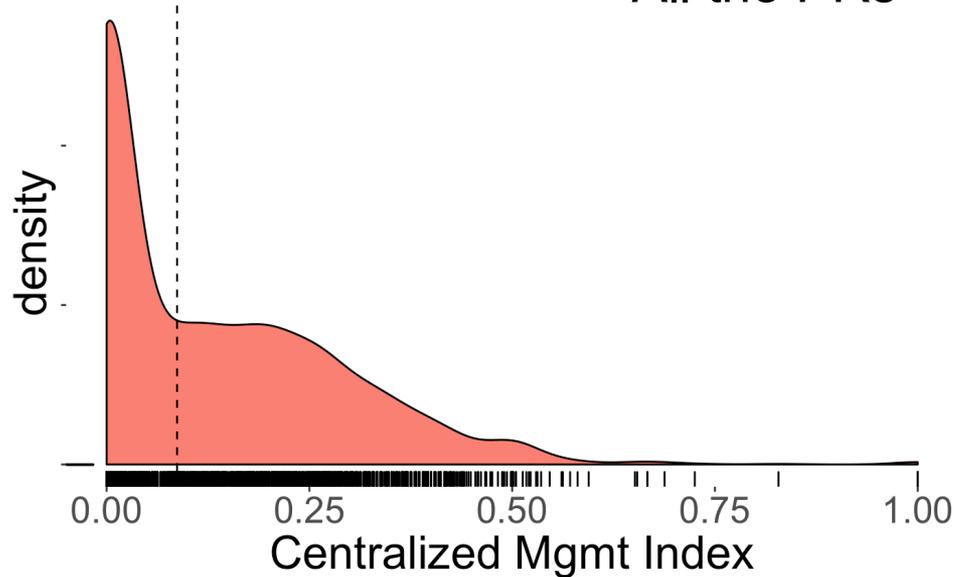
achievermina commented 3 days ago • edited ▾



Issue [#13048](#)

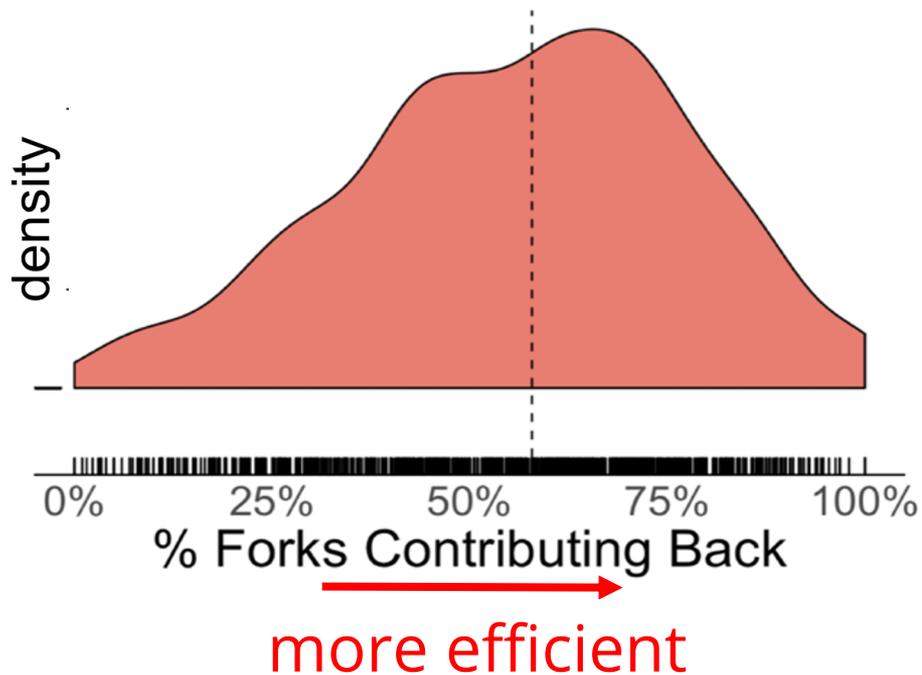
# Operationalization - Centralized Management

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



# Operationalization – Contributing Forks

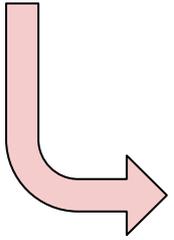
Measure:  $\frac{\text{Number of Forks submitted PR(s)}}{\text{All the Active Forks}}$



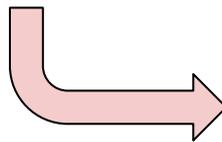
# Test Hypotheses

Group	#fork	#projects	#projects in sample set
A	[3,000 , +]	231	200
B	[1,000 , 3,000)	847	300
C	[20 , 1,000)	116,532	1300

Sampling



Quantifying {  
Inefficiencies  
Practices  
Context Factors



Multiple  
Regression  
Modeling

# Centralized Mgmt → 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

**Trade-offs?**



VS



# Hypotheses

Centralized mgmt → Higher likelihood of community fragm.



Behind the Scenes Bytes

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

by Michael Jones  
Apr 4, 2018

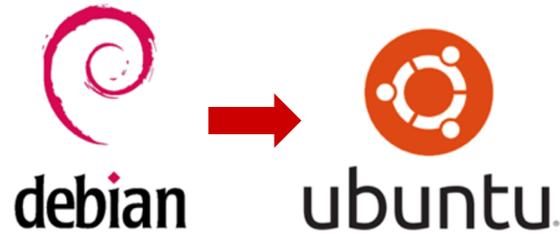


# Old Notion of Forking: Splitting off a Community

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

# Old Notion of Forking: Splitting off a Community

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



# Community Fragn. is Expensive

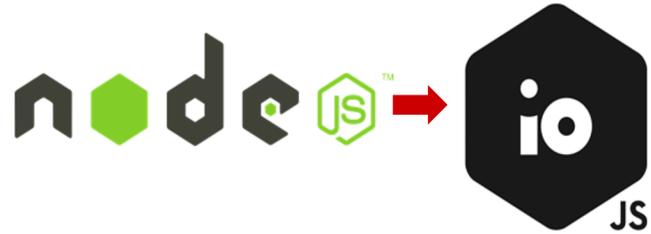


Forking was a Weighty Decision

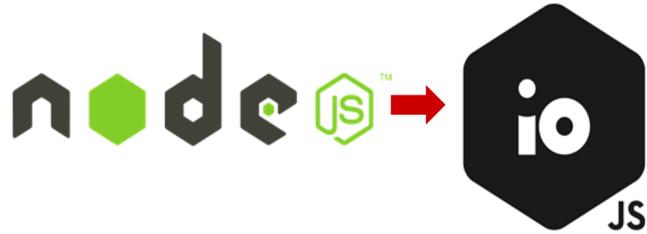
A strong norm against forking [Yoo 2016]



# Community Fragn. is Expensive

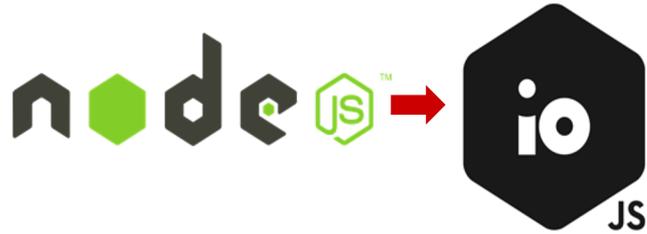


# Community Fragn. is Expensive



“Some open-source forks have *made life difficult for developers.* ... that will force developers to pick sides.” –Lauren Orsini

# Community Fragn. is Expensive



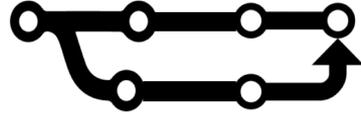
“Some open-source forks have *made life difficult for developers*. ... that will force developers to pick sides.” –Lauren Orsini

**Node.js and io.js are settling their differences, merging back together**



by OWEN WILLIAMS — Jun 16, 2015 in DESIGN & DEV

# Different kinds of Forks



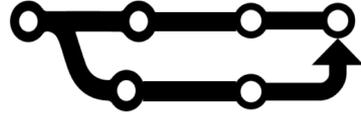
**Hard Fork** **vs** **(Social) Fork**

# Operationalization – Community Fragmentation



**Hard Fork**

**VS**



**(Social) Fork**

# Operationalization – Community Fragmentation



**Hard Fork** **vs** **(Social) Fork**

[ICSE'20]

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

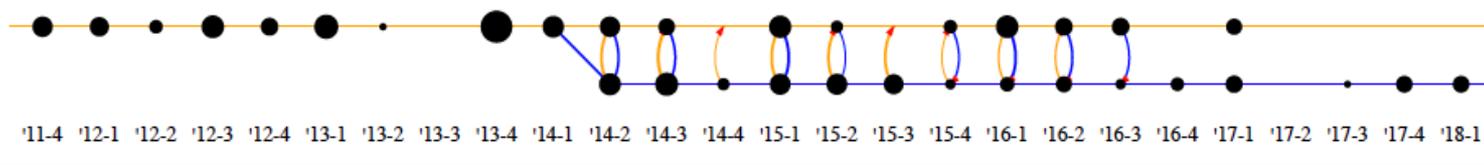
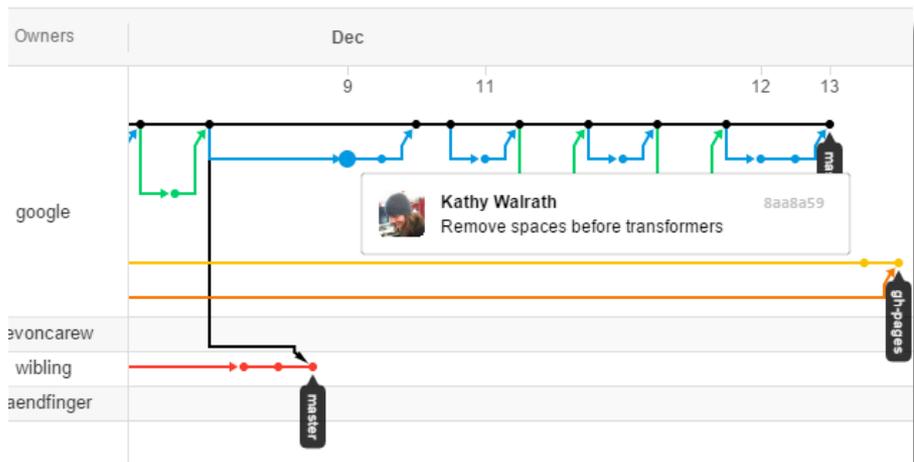
Shurui Zhou  
Carnegie Mellon University, USA

Bogdan Vasilescu  
Carnegie Mellon University, USA

Christian Kästner  
Carnegie Mellon University, USA

# Operationalization – Community Fragmentation

## Detecting Hard Forks



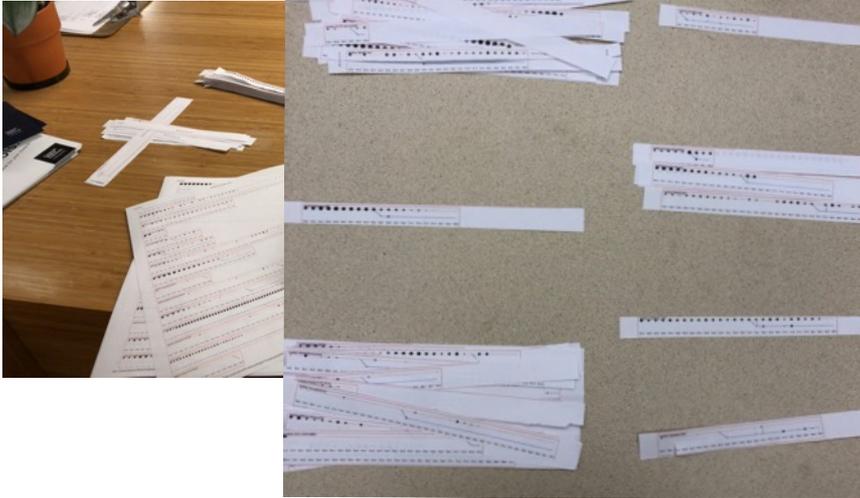
# Operationalization – Community Fragmentation

## Identifying Evolution Patterns of Hard Forks



# Operationalization – Community Fragmentation

## Identifying Evolution Patterns of Hard Forks



# Operationalization – Community Fragmentation

## Identifying Evolution Patterns of Hard Forks

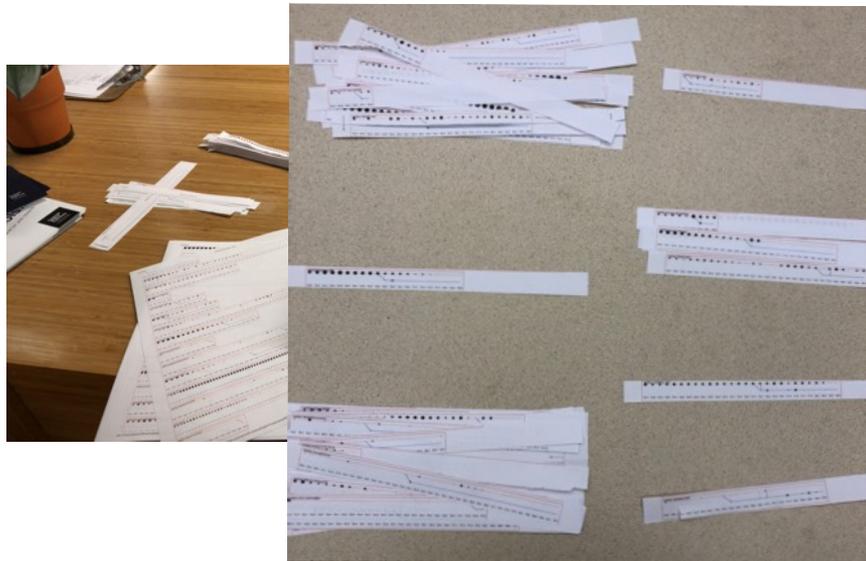


Table 2: Evolution patterns of hard forks

Id	Category	Total	Sub-category	Example	Count	Interviewees	
1	Revive Dead Project	632	Success (F active > 2 Qt)	Upstream remains inactive		576	P12
2			Upstream active again	56			
3			Not success (F active <= 2 Qt)	420			
4	Both Alive	723	only merge		26	P10	
5			only sync		107		P2, P13, P15
6			merge & sync		28		
7	Fork Lived Longer	7280	no interaction		562	P1, P3, P4, P5, P7, P14	
8			only merge		174		
9			only sync		686		
10	Forking Active Project	6251	merge & sync		107	P6, P8, P11	
11			no interaction		6313		
12			only merge		388		
13	Fork does not live upstream	6251	only sync		762	P6, P8, P11	
14			merge & sync		199		
15			no interaction		4902		

# Operationalization – Community Fragmentation

## Identifying Evolution Patterns of Hard Forks

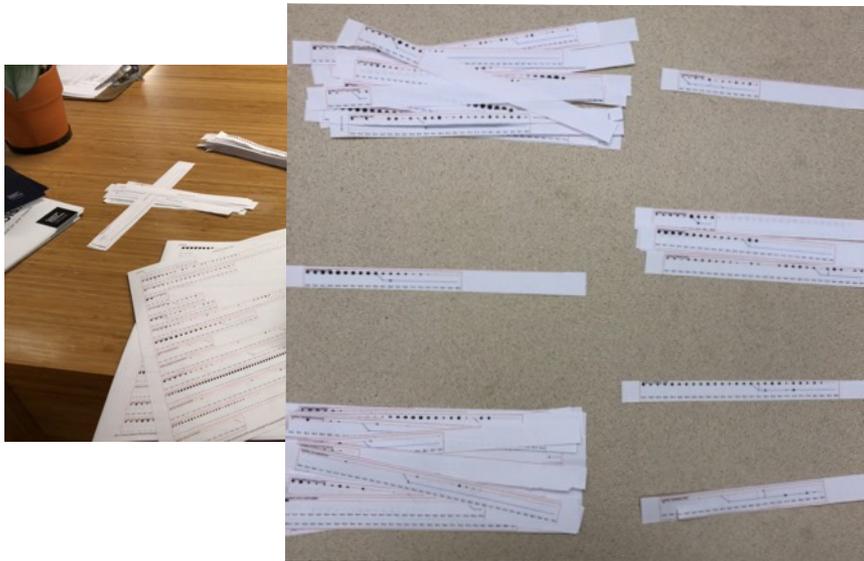


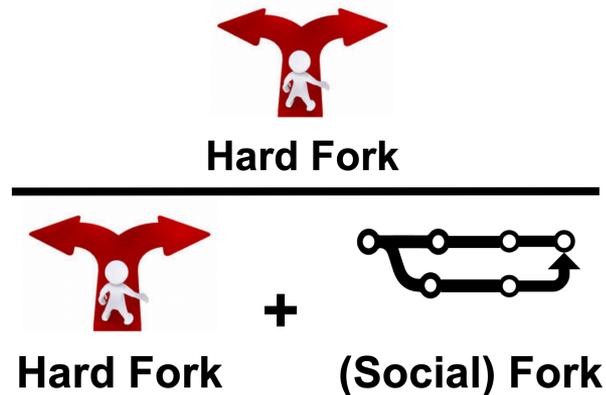
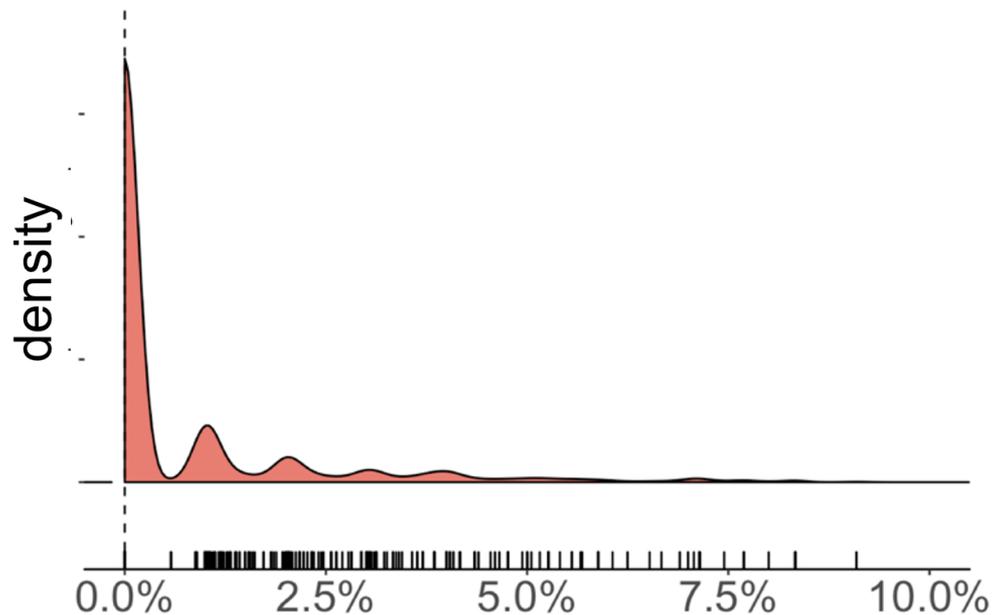
Table 2: Evolution patterns of hard forks

Id	Category	Total	Sub-category	Example	Count	Interviewees
1	Success (F active > 2 Qt)	632	Upstream remains inactive		576	P12
2	Revive Dead Project		Upstream active again		56	
3	Not success (F active <= 2 Qt)	420			420	
4			only merge		26	P10
5	Both Alive	723	only sync		107	P2, P13, P15
6			merge & sync		28	P9
					562	P1, P3, P4, P5, P7, P14
					174	
					686	
					107	
11			no interaction		6313	P6, P8, P11
12			only merge		388	
13	Fork does not out live upstream	6251	only sync		762	
14			merge & sync		199	
15			no interaction		4902	

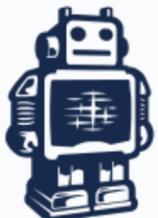


# Operationalization – Community Fragmentation

Ratio of Community Fragmentation:



# Example -- Fragmented Community



Ultimaker

Ultimaker / Ultimaker2Marlin



Hard Fork



MarlinFirmware / Marlin

## add M720 command to list long filenames #118

**Closed** rmd6502 wants to merge 1 commit into `Ultimaker:master` from `rmd6502:M720`

Conversation 1 Commits 1 Checks 0 Files changed 4



rmd6502 commented on **Oct 15, 2016**

Something I found useful - lmk what I need to make it pullable

add M720 command to list long filenames

## [1.1.x] Add 'M27 C' to echo filename (and long name) #10119

**Merged** thinkyhead merged 2 commits into `MarlinFirmware:bugfix-1.1.x` from `thinkyhead:bf1_long_filename_M27` on Mar 16, 2018

Conversation 0 Commits 2 Checks 0 Files changed 3



thinkyhead commented on **Mar 15, 2018**

Member + 😊 ...

Reviewers

No reviews

Based on #10055 by @TheSFReader

# Hypoth: Centralized Mgmt → Community Fragm.

Community Fragmentation



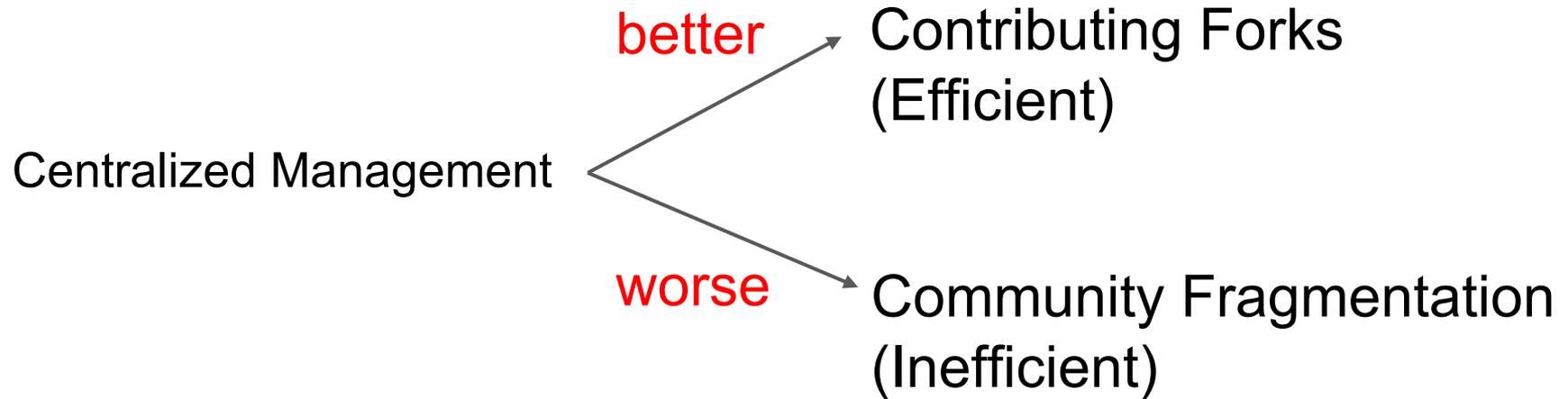
Centralized Management  
(12% of variance explained)

Plus controls for:

Number of Forks

Size

# Trade-off: Centralized Management



# 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



# Research Question

**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

# Solution



Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Natural Intervention

Identifying Best Practices

[FSE'19]

[ICSE'20]

New Intervention

Identifying Feature

[ICSE'18]

Identifying Redundancies

[SANER'19]

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





# Problem

# Solution



Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Natural Intervention

Identifying Best Practices

New Intervention

[FSE'19]

[ICSE'20]

[ICSE'18]

[SANER'19]

- + 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

# Solution



Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Natural Intervention

Identifying Best Practices

New Intervention

Awareness Tools

[FSE'19]

[ICSE'20]

[ICSE'18]

[SANER'19]

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

# Problem

# Solution



Lack of Overview

Lost Contribution

Redundant Development

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

# Solution



Lack of Overview

Lost Contribution

Natural Intervention

Identifying Best Practices

[FSE'19]

[ICSE'20]

New Intervention

Identifying Features

[ICSE'18]

[SANER'19]

- + 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 Waśowski  
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 has un-merged commits**

**Mapping between feature to code changes**



## Email System

- **Signature**
- **Encryption**

<i>Feature id</i>	<i>Keyword List</i>	<i>LOC</i>
<b>Sig.</b>	<b>Signature, isSigned, ...</b>	<b>23</b>
<b>Enc.</b>	<b>Encryption, Decryption, isEncrypted, Decrypt, ...</b>	<b>100</b>



## Email System

- Signature
- Encryption
- Decryption

<i>Feature id</i>	<i>Keyword List</i>	<i>LOC</i>
<b>Sig.</b>	Signature, isSigned, ...	<b>23</b>
<b>Enc.</b>	Encryption, isEncryped, ...	<b>55</b>
<b>Dec</b>	Decryption, Decrypt, ...	<b>45</b>



Feature	Navigation		Keyword List	LOC
<b>Sig.</b>	Prev.	Next	Signature, isSigned, ...	23
<b>Enc.</b>	Prev.	Next	Encryption, Decryption, isEncrypted, Decrypt, ...	100

```

Sig. + connect (ui->MaximumGeneratedBlockWeight, SIGNAL ( textChanged ( const QString &)), this , SLOT ( showR
Sig. + connect (ui->MaximumConsensusBlockWeight, SIGNAL ( textChanged ( const QString &)), this , SLOT ( showR
@@ -207,11 +216,16 @@ void OptionsDialog::setMapper()

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

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

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

# INFOX



Feat		Feature	<a href="#">ibradypod/phantomjs, last commit:May 28</a>	LOC
Sig.		onre.	onresourcerequest, bodi, downloadmultibuffer, qnetworkrepli, respons, qbytearray, data, reply, buffers	28
Enc		hea.	header, getcookiestringfromurl, bodi, cookie, get, qurl, qnetworkrepli, respons, url	10
Sig.		sett.	settings, a, phantomcfg, not, bug, fix, websecurityen, qwebset, setattribute, qwebsettings	2
Sig.		Feature	<a href="#">raff/phantomjs, last commit: Mar 5</a>	LOC
Enc		dow.	download, com, pull, file, ad, support, ariya, http	39
Enc		get	get, qt, are, kei_enter, el, mouse, require, clicks, hard, absolute, setfocus, button, coordinates, keypress	29
Sig.		Feature	<a href="#">ricokahler/phantomjs, last commit:Feb 2</a>	LOC
Sig.		readlin	readlin, asyncreadrequest, asyncread, qobject, qstring, readline, data, qvariant, m_file, m_data, file, read	30
Enc		uint	uint, tmp_value_, value, tmp, octet, qvariant, data, namesize_, readrawdata, fromvalue	80
Enc		frame	frame, bmconsumeok, bmdeliver, method_id, id_enum, bmgetempty, bmreject, bmrecover	29
Enc		cono.	conoack, 0x04, consumeoptions, coexclusive, declar_flag, consumeoption, conolocal, conowait, flag	36
Enc		Feature	<a href="#">DeviaVir/phantomjs, last commit: Jan 25, 2016</a>	LOC
Enc		allow	allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio	7

# INFOX Overview Page

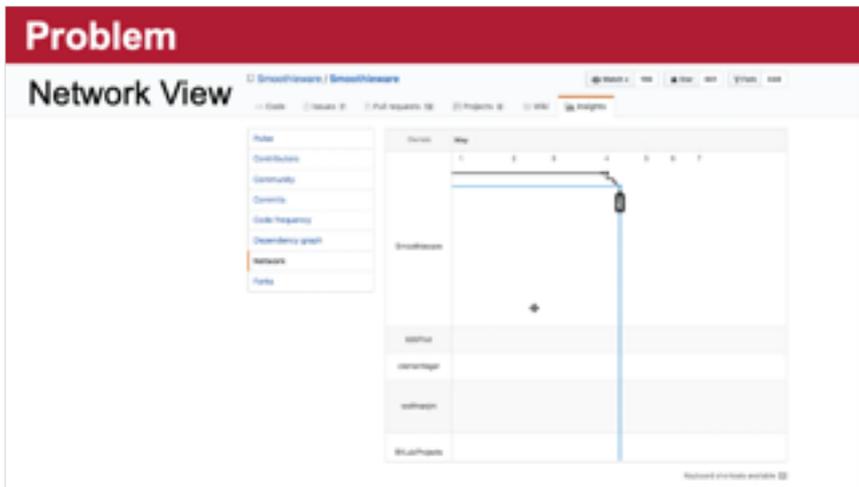


Feature	<a href="#">ibradypod/phantomjs, last commit: May 28</a>	LOC
onre.	onresourcerequest, bodi, downloadmultibuffer, qnetworkrepli, respons, qbytearray, data, reply, buffers	28
hea.	header, getcookiestringfromurl, bodi, cookie, get, qurl, qnetworkrepli, respons, url	10
sett.	settings, a, phantomcfg, not, bug, fix, websecurityen, qwebset, setattrtribute, qwebsettings	2

Feature	<a href="#">raff/phantomjs, last commit: Mar 5</a>	LOC
dow.	download, com, pull, file, ad, support, ariya, http	39
get	get, qt, are, kei_enter, el, mouse, require, clicks, hard, absolute, setfocus, button, coordinates, keypress	29

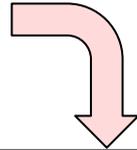
Feature	<a href="#">ricokahler/phantomjs, last commit: Feb 2</a>	LOC
readlin	readlin, asyncreadrequest, asyncread, qobject, qstring, readline, data, qvariant, m_file, m_data, file, read	30
uint	uint, tmp_value_, value, tmp, octet, qvariant, data, namesize_, readrawdata, fromvalue	80
frame	frame, bmconsumeok,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

Feature	<a href="#">DeviaVir/phantomjs, last commit: Jan 25, 2016</a>	LOC
allow	allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio	7

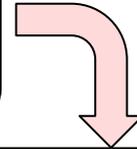




**Dependency graph  
for code changes  
(static analysis)**



**Clustering features  
(community detection)**



**Labeling features  
(NLP)**

# Dependency Graph



## File 1: Email.h

```
1 struct email
2 {
3     char *subject;
4     char *body;
5 + int isEncrypted;
6 };
7 void printMail ( struct email *msg);
8
9 + int isEncrypted (struct email *msg);
10
11 + int isSigned (struct email *msg);
```

## File 2: Email.c

```
1 + void printMail ( struct email *msg)
2 {
3     printf ("SUBJECT:", msg->subject );
4 + printf ("SIGNED:", msg->isSigned);
5 + if (0 == (isEncrypted(msg) ))
6         printf ( "BODY:", msg->body );
7 + else
8 +     printf ( "Encrypted msg." );
9 }
10
11 + int isEncrypted (struct email *msg)
12 + {
13 +     return msg->isEncrypted;
14 + }
15
16 + int isSigned (struct email *msg)
17 + {
18 +     return msg->isSigned;
19 + }
```

### 3 Dependencies

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

# Dependency Graph



## File 1: Email.h

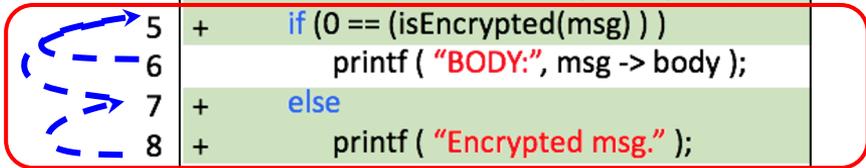
```
1 struct email
2 {
3     char *subject;
4     char *body;
5 + int isEncrypted;
6 };
7 void printMail ( struct email *msg);
8
9 + int isEncrypted (struct email *msg);
10
11 + int isSigned (struct email *msg);
```

## 3 Dependencies

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

## File 2: Email.c

```
1 + void printMail ( struct email *msg)
2 {
3     printf ("SUBJECT:", msg -> subject );
4 + printf ("SIGNED:", msg->isSigned);
5 + if (0 == (isEncrypted(msg) ) )
6     printf ( "BODY:", msg -> body );
7 + else
8 +     printf ( "Encrypted msg." );
9 }
10
11 + int isEncrypted (struct email *msg)
12 + {
13 +     return msg->isEncrypted;
14 + }
15
16 + int isSigned (struct email *msg)
17 + {
18 +     return msg->isSigned;
19 + }
```



# Dependency Graph



## File 1: Email.h

```
1 struct email
2 {
3     char *subject;
4     char *body;
5 + int isEncrypted;
6 };
7 void printMail ( struct email *msg);
8
9 + int isEncrypted (struct email *msg);
10
11 + int isSigned (struct email *msg);
```

## 3 Dependencies

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

## File 2: Email.c

```
1 + void printMail ( struct email *msg)
2 {
3     printf ("SUBJECT:", msg -> subject );
4 + printf ("SIGNED:", msg->isSigned);
5 + if (0 == (isEncrypted(msg) ) )
6         printf ( "BODY:", msg -> body );
7 + else
8 +     printf ( "Encrypted msg." );
9     }
10
11 + int isEncrypted (struct email *msg)
12 + {
13 +     return msg->isEncrypted;
14 + }
15
16 + int isSigned (struct email *msg)
17 + {
18 +     return msg->isSigned;
19 + }
```

# Dependency Graph



## File 1: Email.h

```
1 struct email
2 {
3     char *subject;
4     char *body;
5 + int isEncrypted;
6 };
7 void printMail ( struct email *msg);
8
9 + int isEncrypted (struct email *msg);
10
11 + int isSigned (struct email *msg);
```

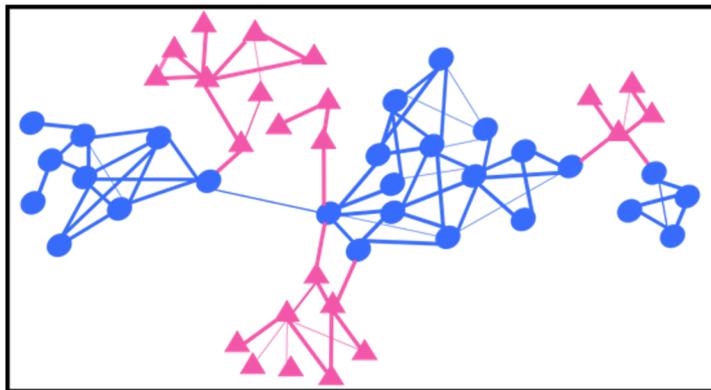
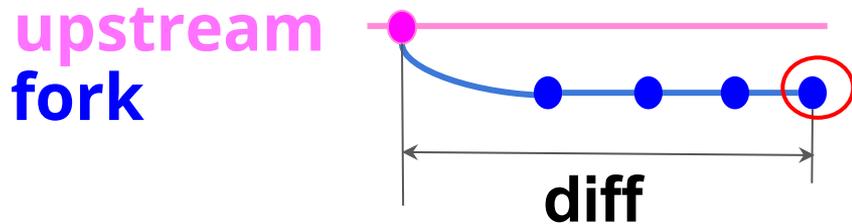
## 3 Dependencies

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

## File 2: Email.c

```
1 + void printMail ( struct email *msg)
2 {
3     printf ("SUBJECT:", msg -> subject );
4 + printf ("SIGNED:", msg->isSigned);
5 + if (0 == (isEncrypted(msg) ) )
6         printf ( "BODY:", msg -> body );
7 + else
8 +     printf ( "Encrypted msg." );
9     }
10
11 + int isEncrypted (struct email *msg)
12 + {
13 +     return msg->isEncrypted;
14 + }
15
16 + int isSigned (struct email *msg)
17 + {
18 +     return msg->isSigned;
19 + }
```

# Dependency Graph



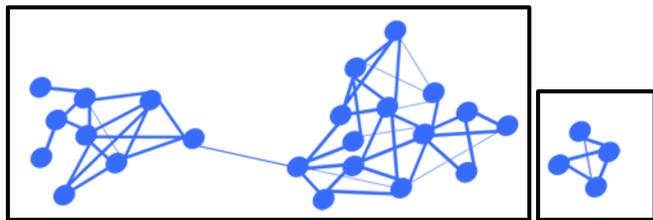
- labeled, changed code
- ▲ base code

Dependency  
graph

Clustering  
features

Labeling  
features

# Dependency Graph



**Dependency graph**

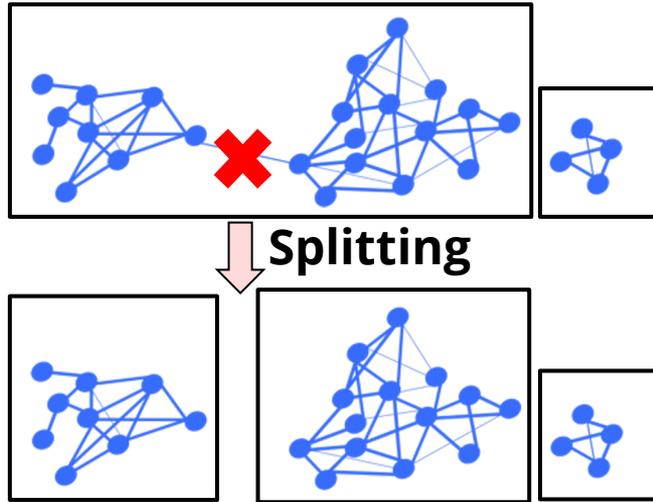


**Clustering features**



**Labeling features**

# Dependency Graph

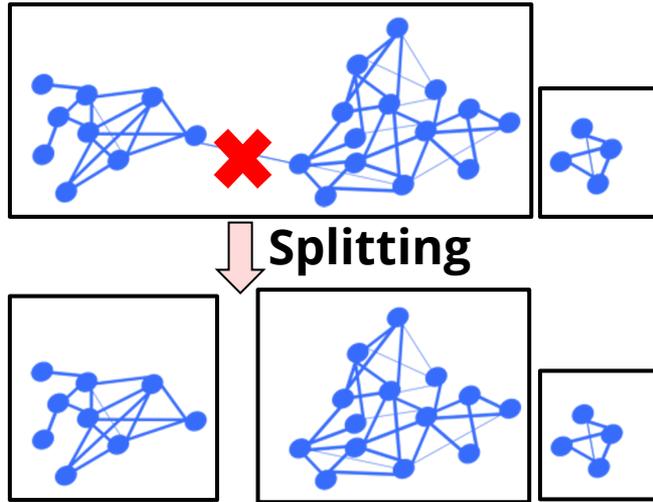


**Dependency graph**

**Clustering features**

**Labeling features**

# Dependency Graph



Network Analysis  
Girvan–Newman algorithm

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

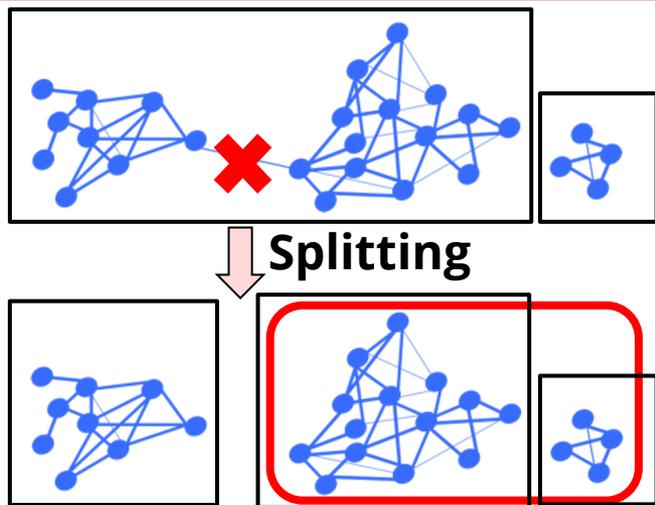


Dependency graph

Clustering features

Labeling features

# Dependency Graph

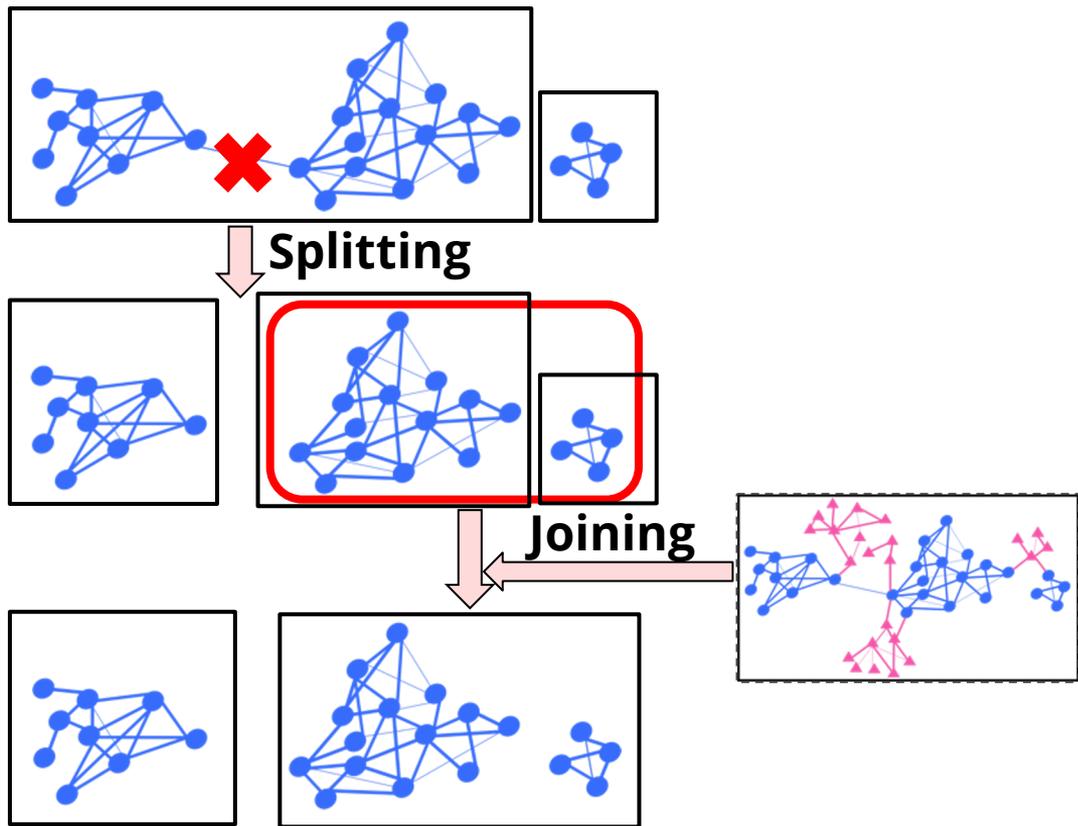


**Dependency graph**

**Clustering features**

**Labeling features**

# Dependency Graph

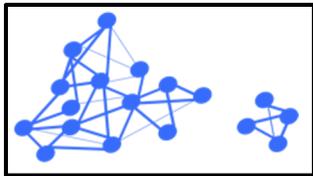
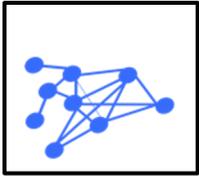


Dependency graph

Clustering features

Labeling features

# Dependency Graph



- **commit message**
- **code**
- **comment**

↓ **TF-IDF, N-Gram**

<b>Sig.</b>	Signature, isSigned, ...	23
<b>Enc.</b>	Encryption, Decryption, isEncrpyed, Decrypt, ...	100

**Dependency graph**



**Clustering features**



**Labeling features**

# INFOX



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



Feature	<a href="#">raff/phantomjs, last commit: Mar 5</a>	LOC
dow.	download, com, pull, file, ad, support, ariya, http	39
get	get, qt, are, kei_enter, el, mouse, require, clicks, hard, absolute, setfocus, button, coordinates, keypress	29



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



Feature	<a href="#">DeviaVir/phantomjs, last commit: Jan 25, 2016</a>	LOC
allow	allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio	7



Feature	Navigation		Keyword List	LOC
<b>Sig.</b>	Prev.	Next	Signature, isSigned, ...	23
<b>Enc.</b>	Prev.	Next	Encryption, Decryption, isEncrypted, Decrypt, ...	100

```

Sig. + connect (ui->MaximumGeneratedBlockWeight, SIGNAL ( textChanged ( const QString &)), this , SLOT ( showR
Sig. + connect (ui->MaximumConsensusBlockWeight, SIGNAL ( textChanged ( const QString &)), this , SLOT ( showR
@@ -207,11 +216,16 @@ void OptionsDialog::setMapper()

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

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

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

```



## 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?**



**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**

Project	#Features
Cherokee	328
clamav	285
ghostscript	816
Marlin	280
MPSolve	17
openvpn	276
subversion	409
tcl	2,481
xorg-server	1,360
xterm	453

# INFOX - Effectiveness



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

**INFOX assigned features  
with 90% accuracy.**

Project	#Features
Cherokee	328
clamav	285
ghostscript	816
Marlin	280
MPSolve	17
openvpn	276
subversion	409
tcl	2,481
xorg-server	1,360
xterm	453

# INFOX - Evaluation



**Effectiveness**

**Usefulness**

**Quantitative  
Study**

**Human-subject Study**



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



# Human-subject Study



- **11 developers**



Project	#Forks
MarlinFirmware/Marlin	4,149
Smoothieware/Smoothieware	566
grpc/grpc	2,226
timscaffidi/ofxVideorecorder	60
arduino/Arduino	5,592
bitcoin/bitcoin	9,696
ariya/phantomjs	4,921

# Human-subject Study - Effectiveness



**Effectiveness**

**Usefulness**

**Quantitative  
Study**

**Human-subject Study**

**Most of the developers agree with the features that INFOX detected after a few steps of splitting and merging.**

# Human-subject Study - Usefulness



## Can INFOX help developers to gain a better overview of repository forks?

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

Feature	<a href="#">raff/phantomjs, last commit: Mar 5</a>	LOC
dow.	download, com, pull, file, ad, support, ariya, http	39
get	get, qt, are, kei_enter, el, mouse, require, clicks, hard, absolute, setfocus, button, coordinates, keypress	29

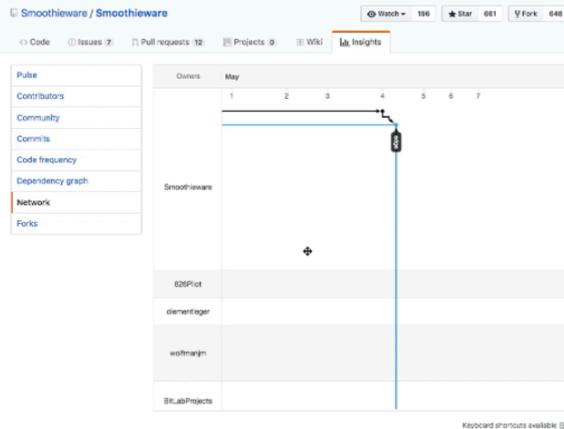
Feature	<a href="#">ricokahier/phantomjs, last commit: Feb 2</a>	LOC
readlin	readlin, asyncreadrequest, asyncread, qobject, qstring, readline, data, qvariant, m_file, m_data, file, read	30
uint	uint, tmp_value_, value, tmp, octet, qvariant, data, namesize_, readrawdata, fromvalue	80
frame	frame, bmconsumeok,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

Feature	<a href="#">DeviaVir/phantomjs, last commit: Jan 25, 2016</a>	LOC
allow	allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio	7

VS

### Problem

### Network View





## Interesting and Reusable Contribution

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

# Human-subject Study - Usefulness



## 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.”*

forked from MarlinFirmware/Marlin

<> Code Pull requests 0 Projects 0

**Added laser controls to main buffer**  
Faster processing and no laser delays

Marlin\_v1

committed on Jan 4, 2014

forked from MarlinFirmware/Marlin

<> Code Pull requests 0 Projects 0

**Add laser control**  
1.1.x

committed on May 23, 2017



Fork 8.6k



[Followed repositories](#)

[Import your repository](#)

[Search on GitHub](#)

[About us](#)

## Atom/atom

:atom: The hackable text editor

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

[Copy](#) [CSV](#) [Excel](#) [Print](#)

Show  entries

Fork	Commits	Changed files	Lines of code changed	Representative Keyword	Last Commit	Create	Add Tags
<a href="#">ToniFerra72/atom</a>	1	1	2	hola	2018-02-22	2018-02-22	
<a href="#">larsdroid/atom</a>	2	1	1	developing, documented, api, reference, docs, for	2018-02-14	2018-02-13	
<a href="#">alexheretic/atom</a>	2	1	3	normalizedpath, oldpath, normalized, startswith, starts, old	2018-01-24	2018-01-24	
<a href="#">rk162/atom</a>	1	74	2795	class, div, href, img, compatible, jpg	2018-01-12	2018-01-12	
<a href="#">elusora/atom</a>	1	1	1	transparent	2018-01-10	2018-01-10	

Showing 1 to 5 of 106 entries

# Problem

# Solution



Lack of Overview

Redundant Development

Natural Intervention

Identifying Best Practices

New Intervention

Identifying Redundancies

[FSE'19]

[ICSE'20]

[ICSE'18]

[SANER'19]

- + 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 Wasowski  
IT University of Copenhagen, Denmark

# Problem -- Redundant Development



foosel commented on Aug 22, 2017

Owner



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

Member



Duplicate of [#5869](#) and [#5972](#), partially also [#5879](#).

# Cost / Waste

## **For maintainer:**

- **Maintenance effort**

  - Before a duplicate PR is identified:

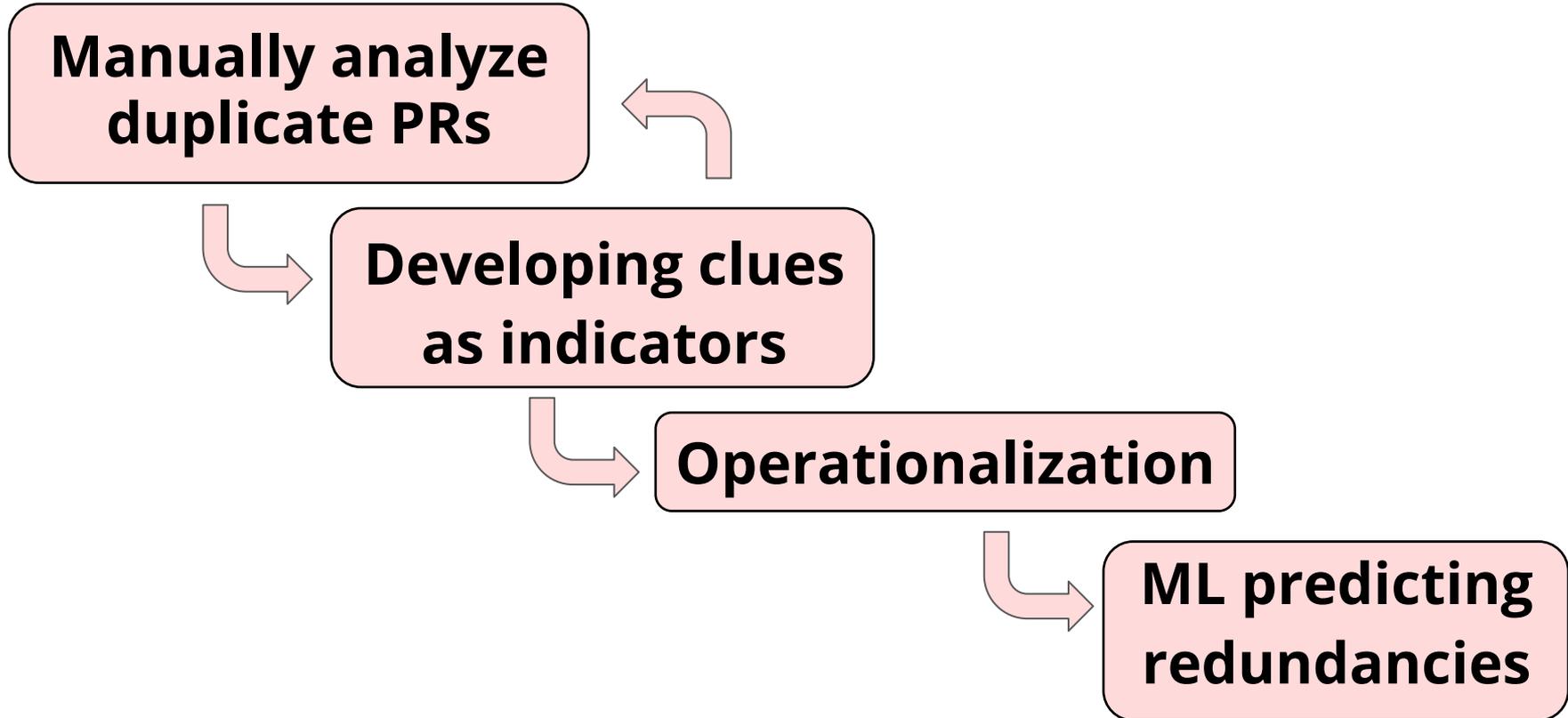
    - 2.6 reviewers

    - 5.2 review comments [Li et al. 2018]

## **For developers:**

- **De-motivate developers** [Steinmacher et al. 2018]

# Research Method

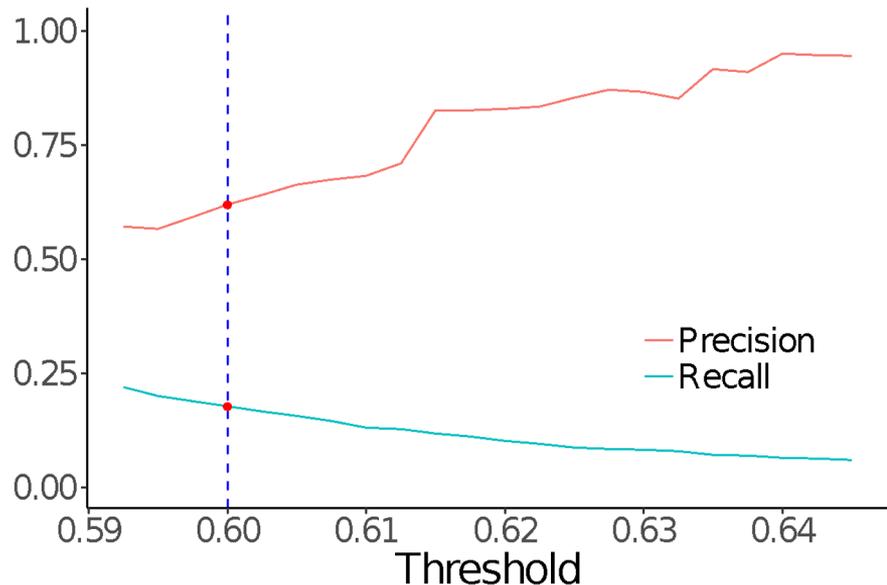


# Evaluation - Effectiveness

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?

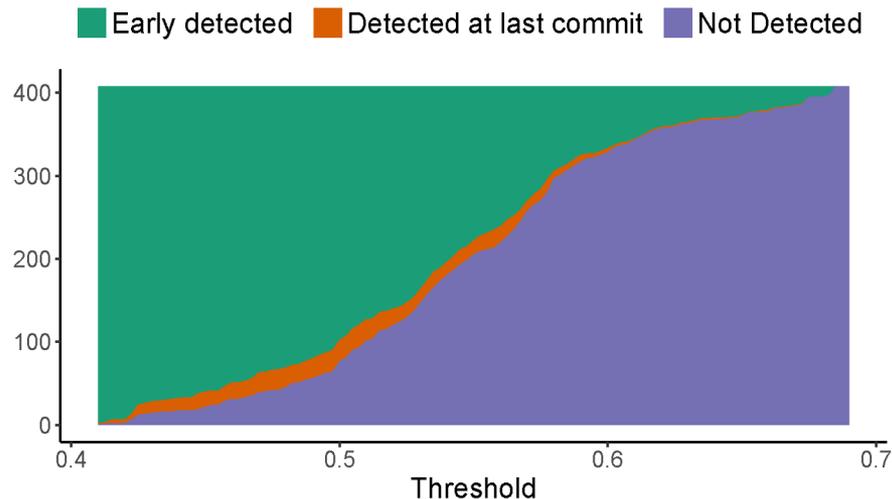
# RQ1: helping maintainers to find duplicate PRs



Randomly sample 400 PRs from each project

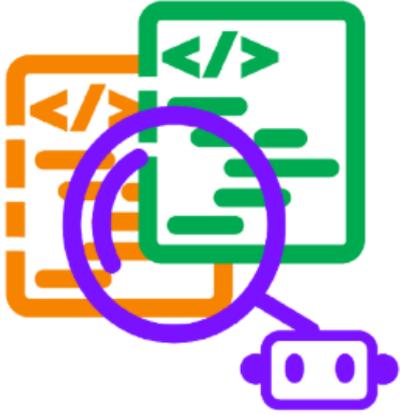
Precision 55%-82%  
Recall 10%-25%

## RQ2: helping developers to find duplicate changes early



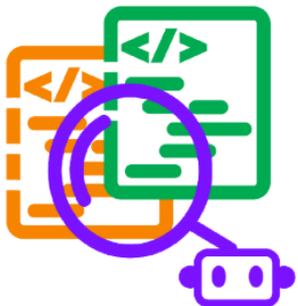
Recall 46% - 71%  
0.07–0.5% false positive rate  
Save 1.9 - 3.0 commits per PR

# Application Scenario



**DuplicatePR-bot**

# Application Scenario



## DuplicatePR-bot



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

Author + 😊 ...

Duplicate [#1370](#), close

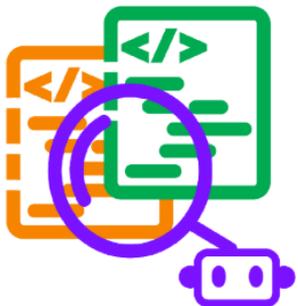


1



sergeyrolich closed this 19 days ago

# Application Scenario



## DuplicatePR-bot



DuplicatePR-bot commented 19 days ago

Hi there! This pull request looks like it might be a duplicate of [#1370](#), *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** ne
- This pull request is **not a duplicate**, so this comment was **not useful**
- 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 mos  
If you are interested in using our bot in the future, please [subscribe](#). If  
see our [web page](#).



sergeyrolich commented 19 days ago

Duplicate [#1370](#), close



sergeyrolich closed this 19 days ago

## Duplicate PR pairs successfully detected:

repo	PR1	PR2
gomods/athens	1423	1370
matomo-org/matomo	14137	14975
mrdoob/three.js	10597	13706
mrdoob/three.js	6230	6301
spring-projects/spring-framework	1695	1817
spring-projects/spring-framework	1492	1628
spring-projects/spring-framework	1218	1653
spring-projects/spring-framework	569	566
spring-projects/spring-framework	758	1827
hashicorp/terraform	18186	18296
cocos2d/cocos2d-x	14883	13687
cocos2d/cocos2d-x	14794	7565
facebook/react	12760	13169

# Problem

# Solution



Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

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



# Limitation

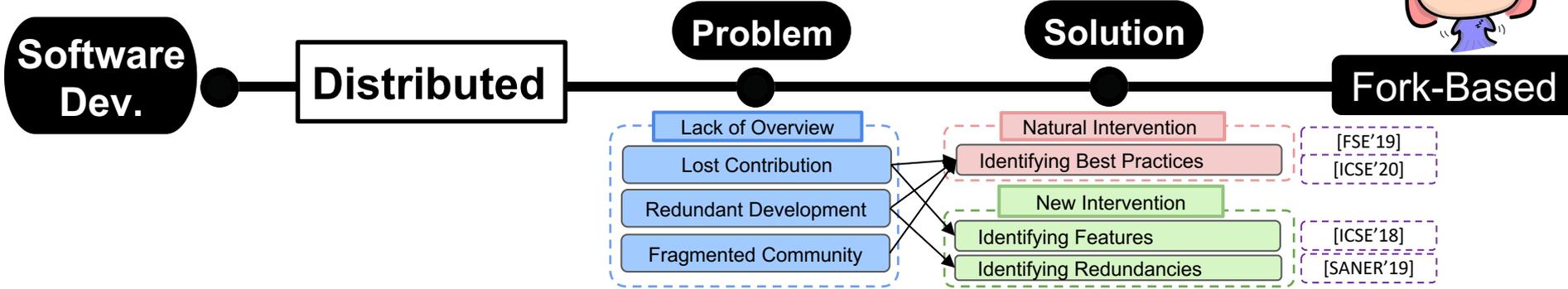
Generalizability

# Limitation

Generalizability

Construct Validity

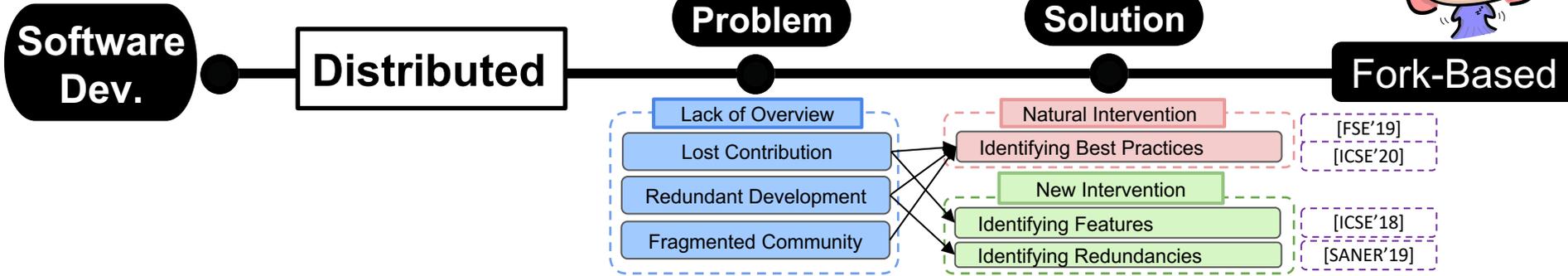
# Improving Collaboration Efficiency



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



# Improving Collaboration Efficiency



**FUTURE**

**Centralized Management Index 98%**

**Modularity High**

**Overview of Hard Forks**

# Improving Collaboration Efficiency



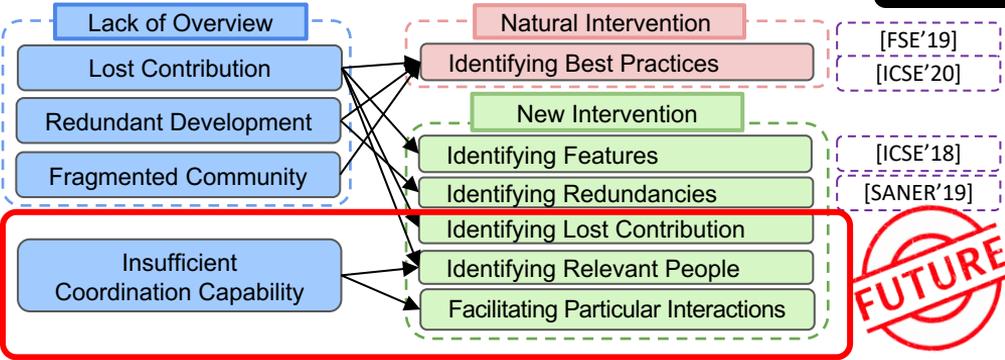
Software Dev.

Distributed

Problem

Solution

Fork-Based



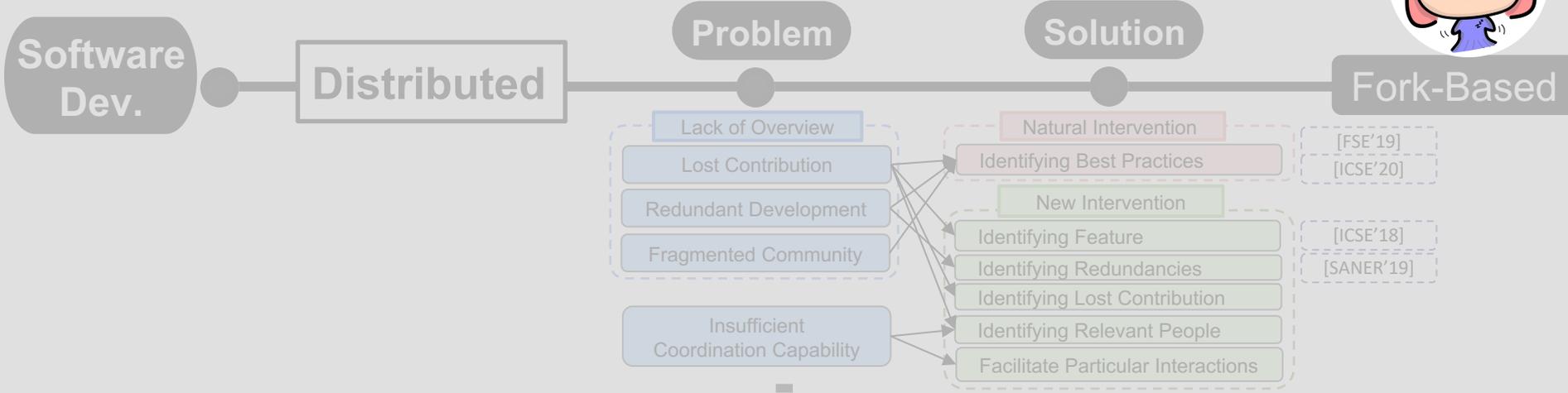
Centralized Management Index 98%

Modularity High

Overview of Hard Forks



# Improving Collaboration Efficiency



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





# Improving Collaboration Efficiency

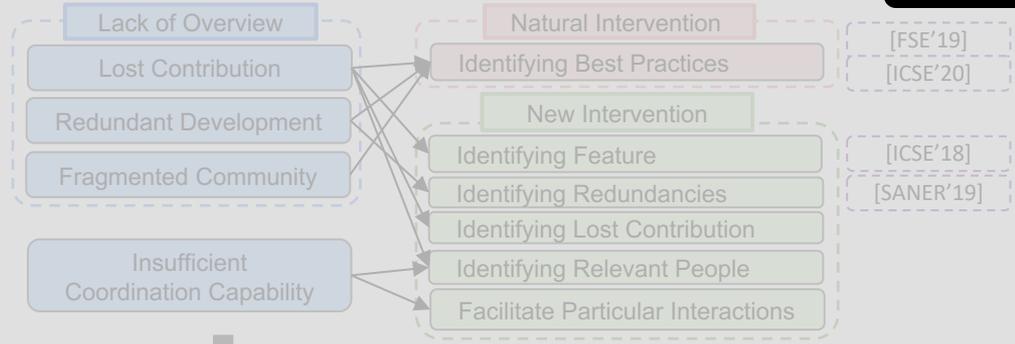
Software Dev.

Distributed

Problem

Solution

Fork-Based



Interdisciplinary

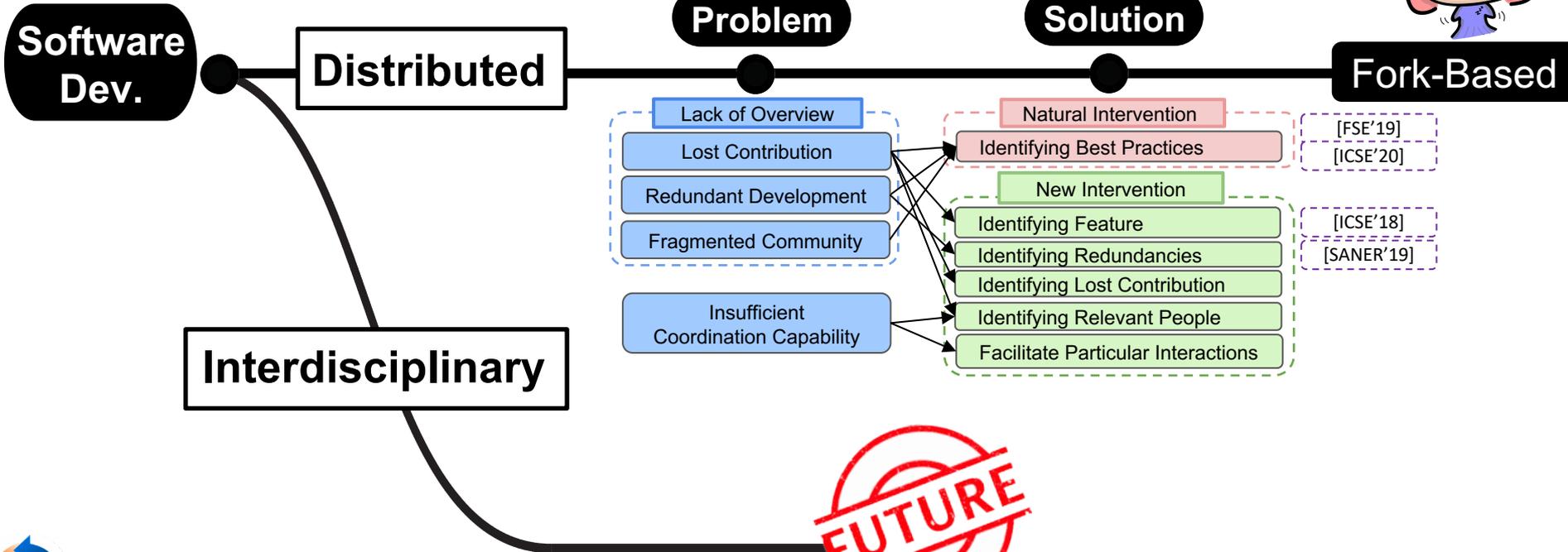


- + Advances in tooling & SE principles
- + Insights from other disciplines
- + Mix a wide range of research methods





# Improving Collaboration Efficiency



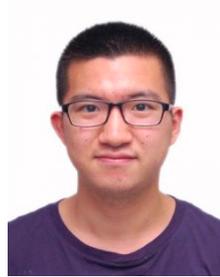
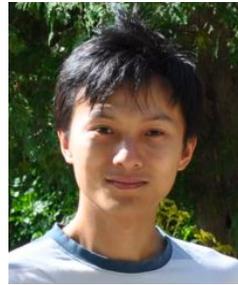
**FUTURE**



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



# Acknowledgement



It's not yet the end ...