Improving Collaboration Efficiency in Fork-based Development

Shurui Zhou

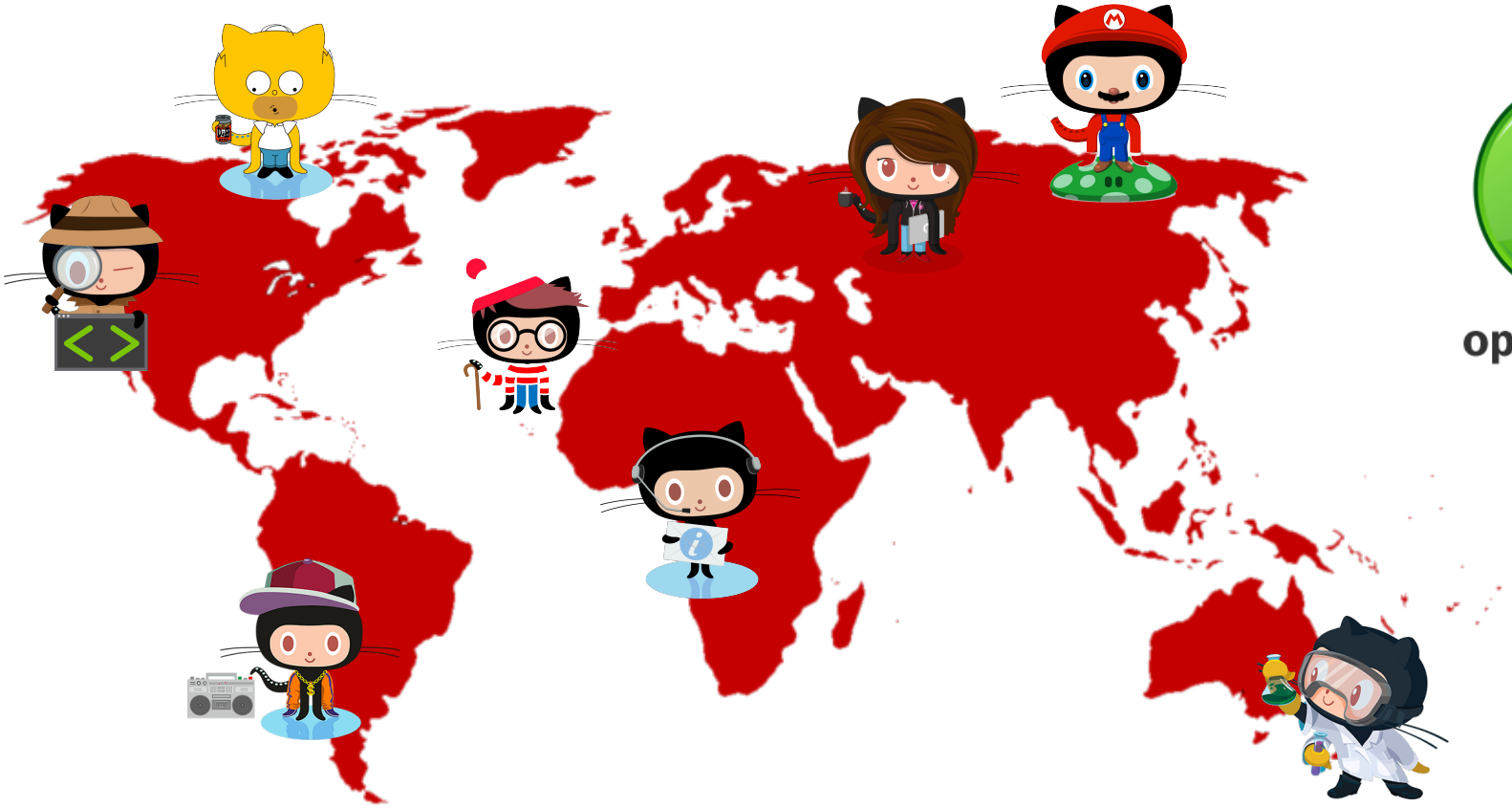
Thesis Committee: Christian Kästner (Chair), James D. Herbsleb, Laura A. Dabbish, Andrzej Wąsowski

Carnegie Mellon University

Collaboration is Everywhere

Globally Distributed Software Development

Globally Distributed Software Development



open source

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

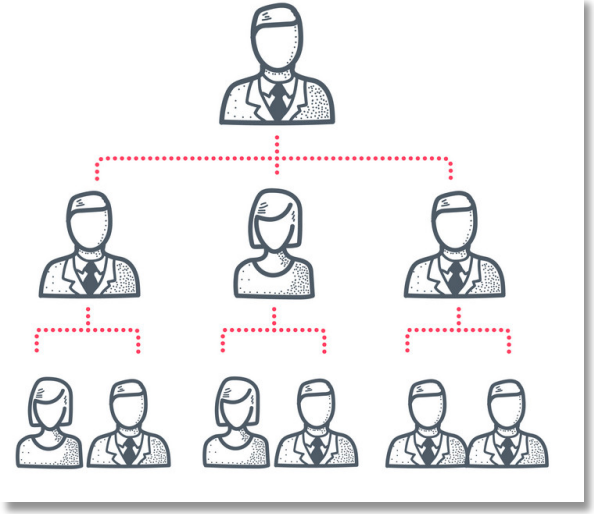
Problem → Intervention → Evaluation

Help software developers
to better collaborate

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

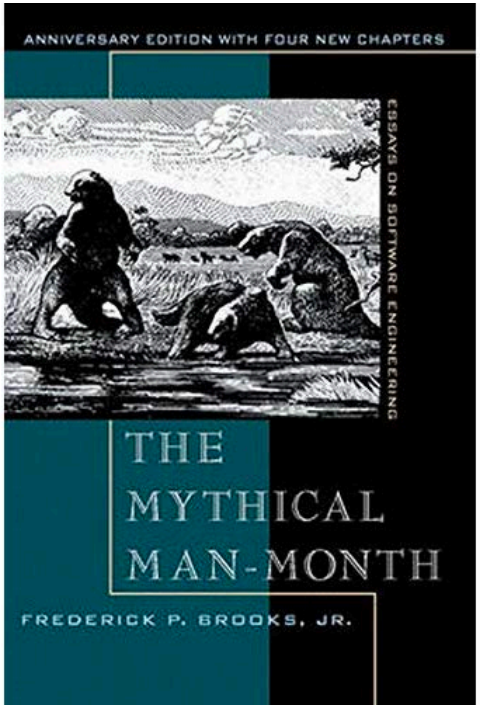
Traditional Collaboration Model

Traditional Collaboration Model

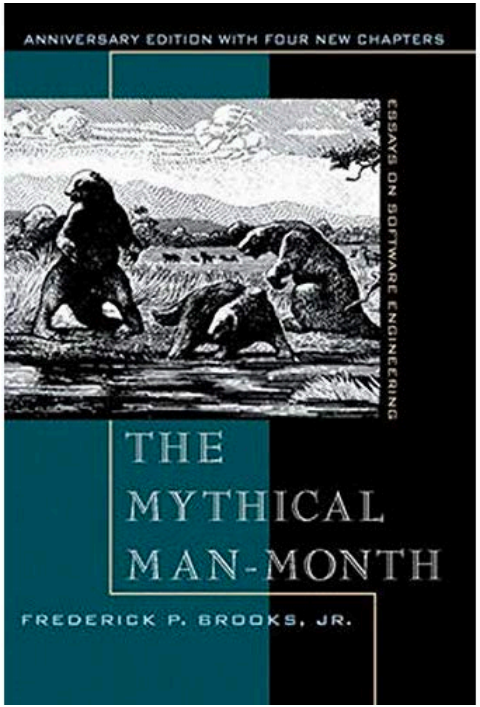


Traditional Collaboration Model

Traditional Collaboration Model



Traditional Collaboration Model



CSCW

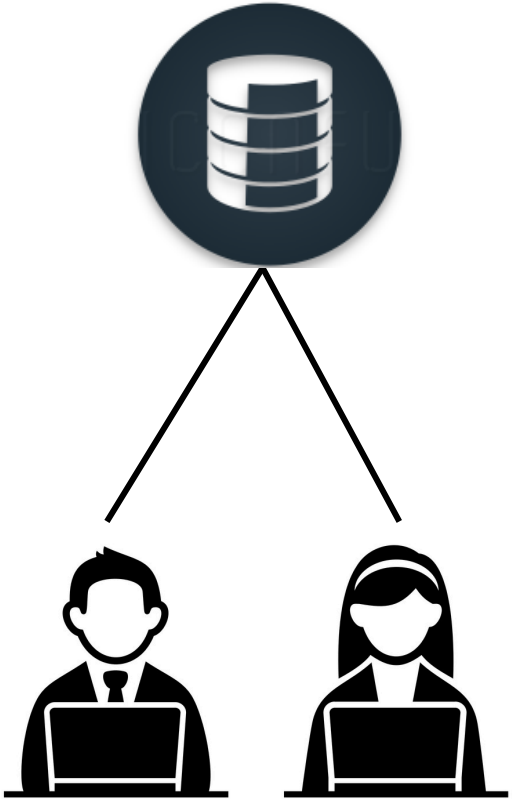
Computer Supported Cooperative Work

Fork-Based Development Changed Everything

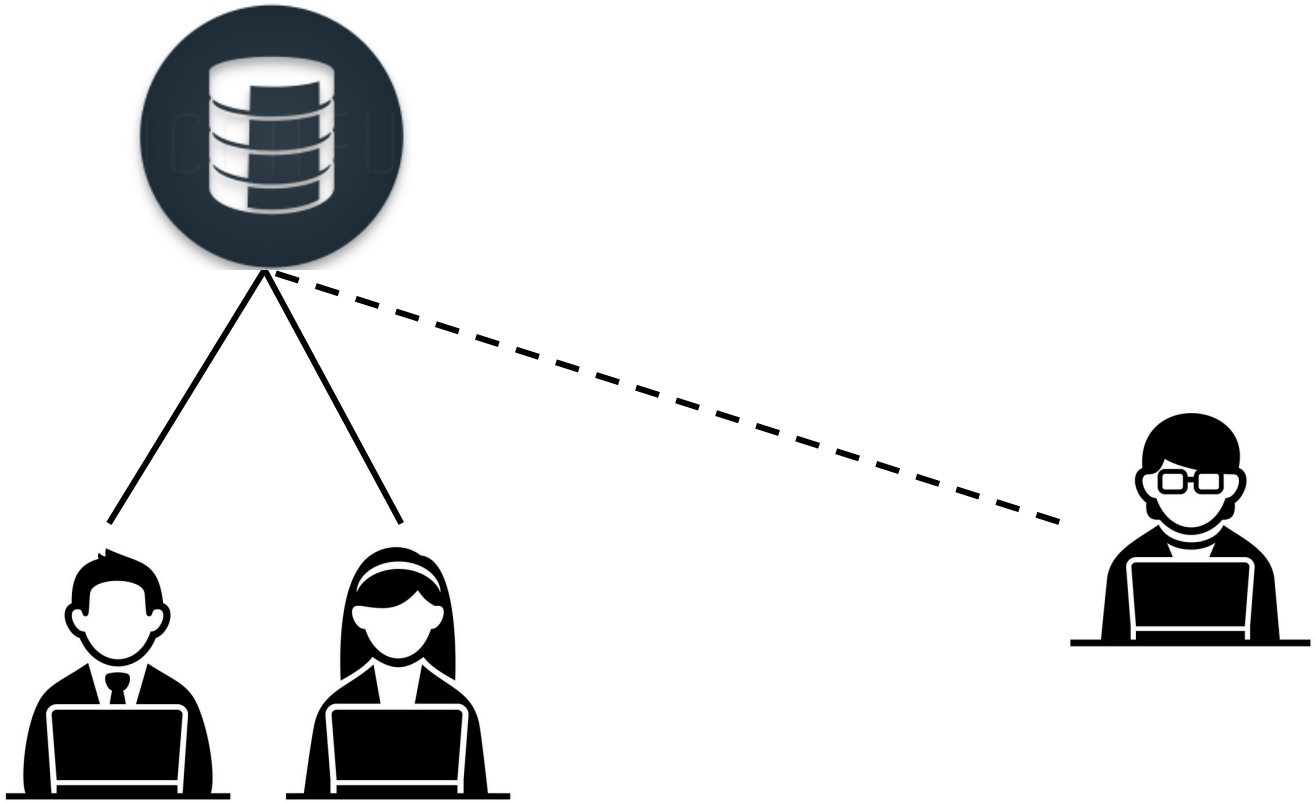
Fork-based Dev. Changed Everything

Traditional Collaboration Model

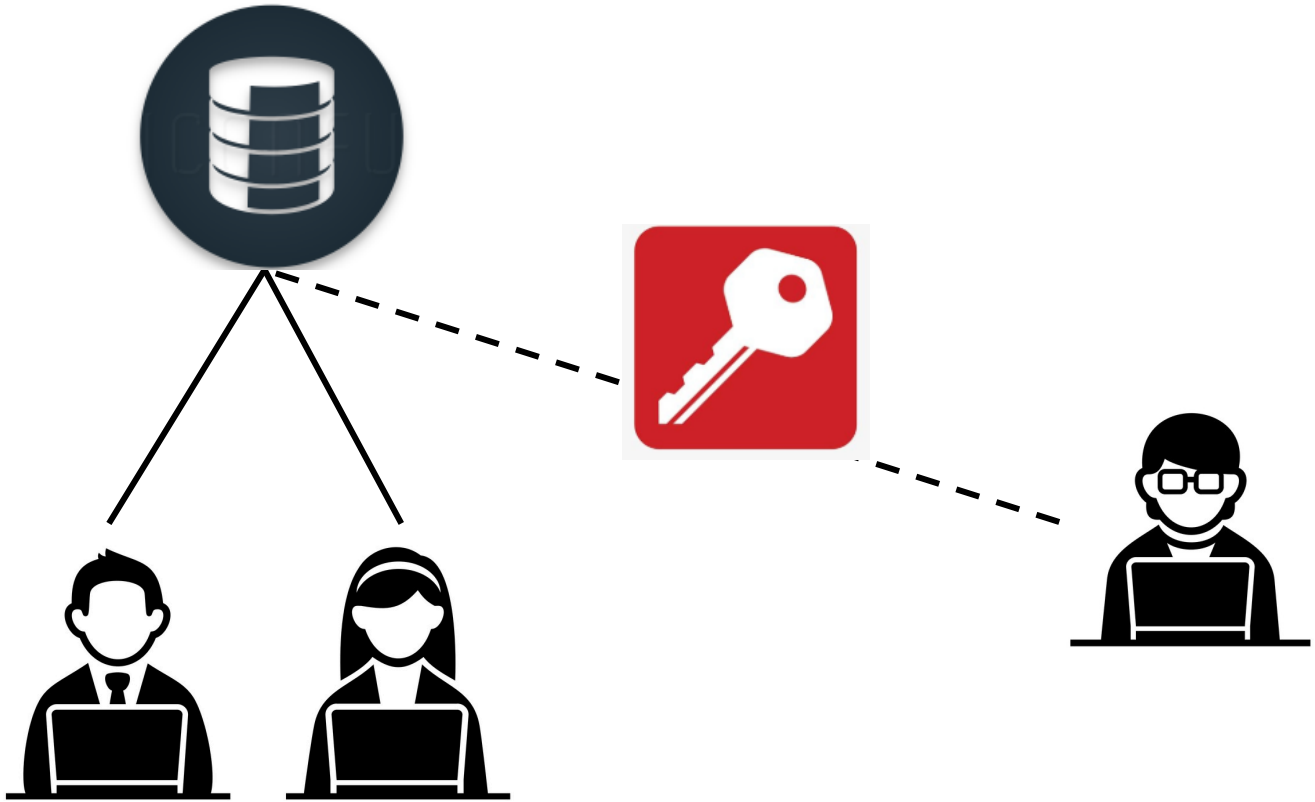
Traditional Collaboration Model



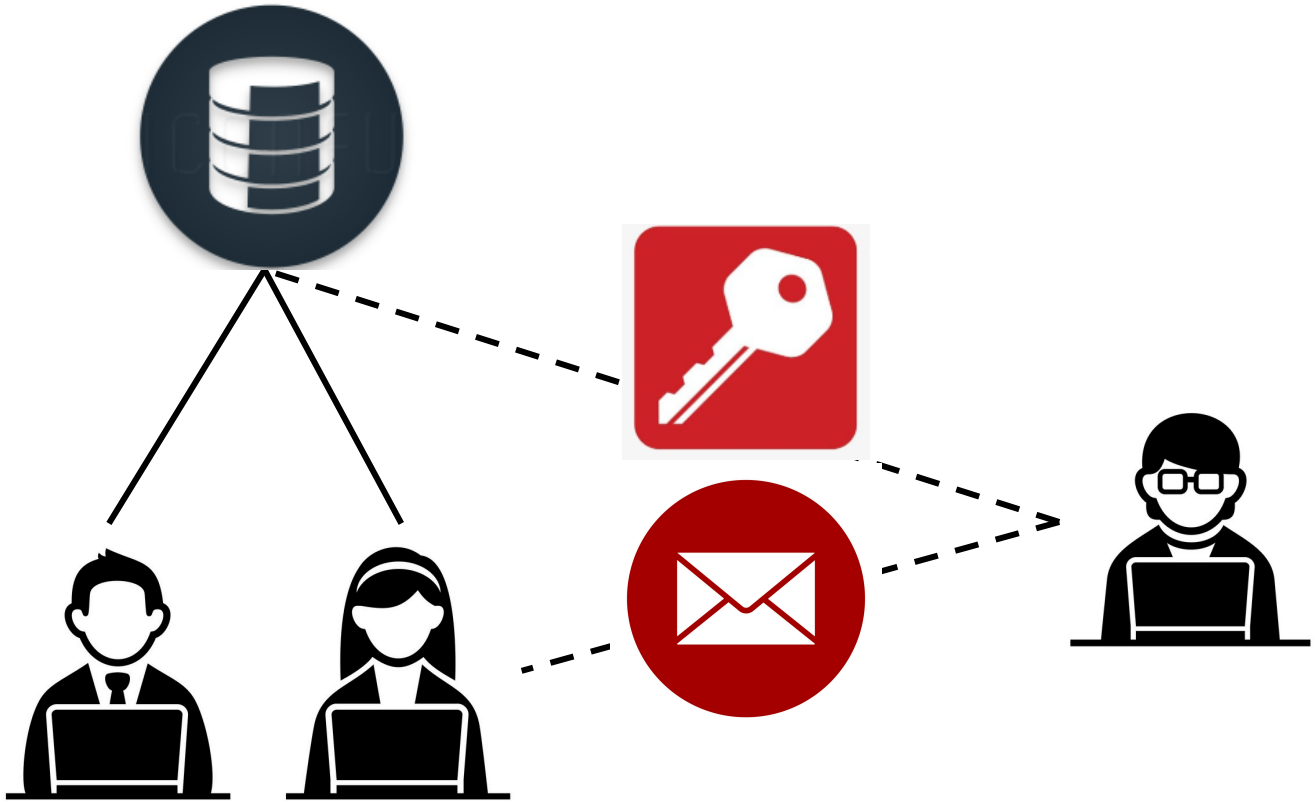
Traditional Collaboration Model



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
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
```

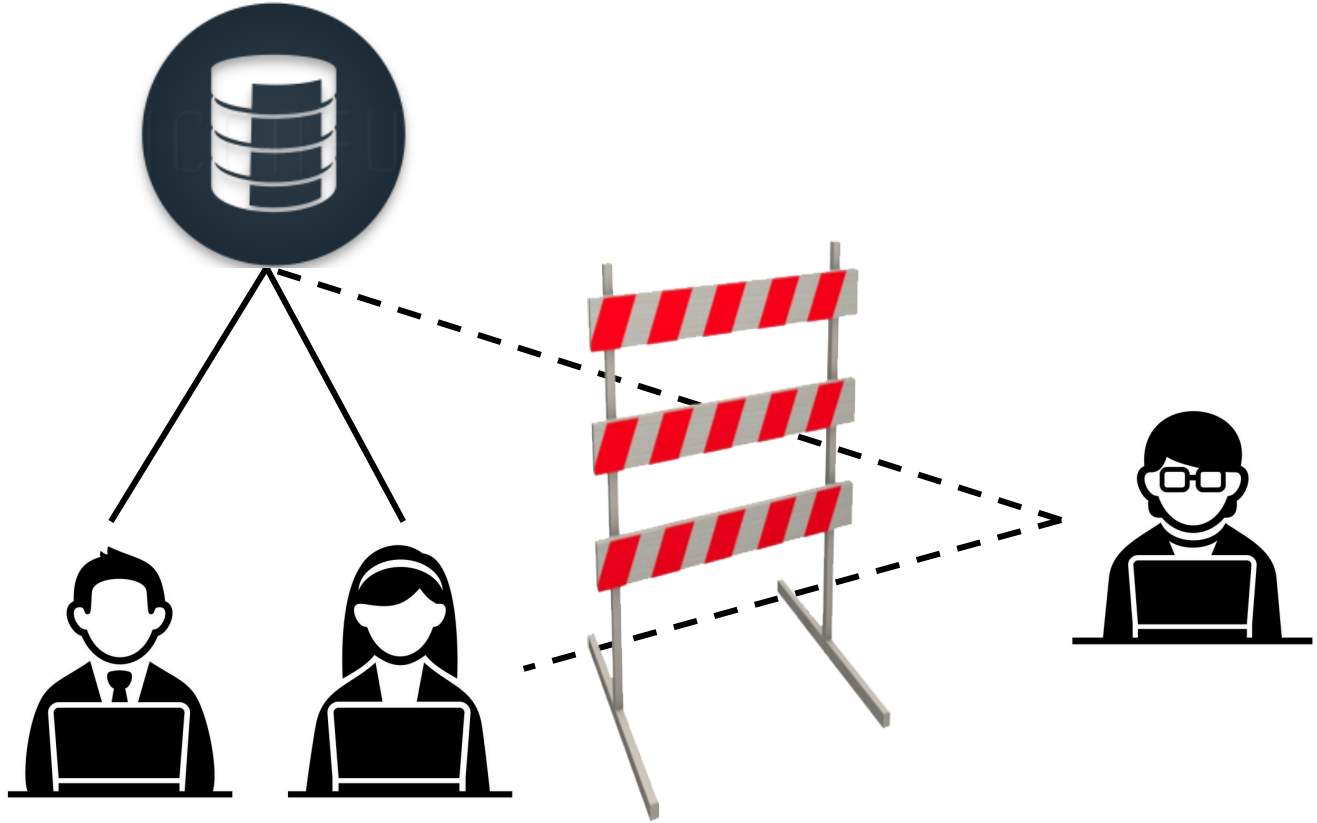
Source code

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

--------------------------------------------------------------------------------
diff -u -ruN 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 kerneld.) 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

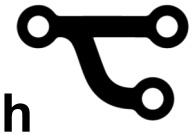
open source



Fork-based Development

Upstream

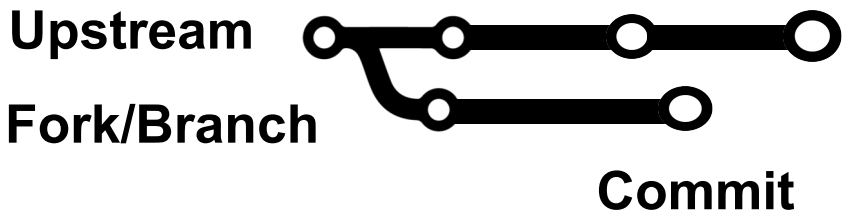
Fork/Branch



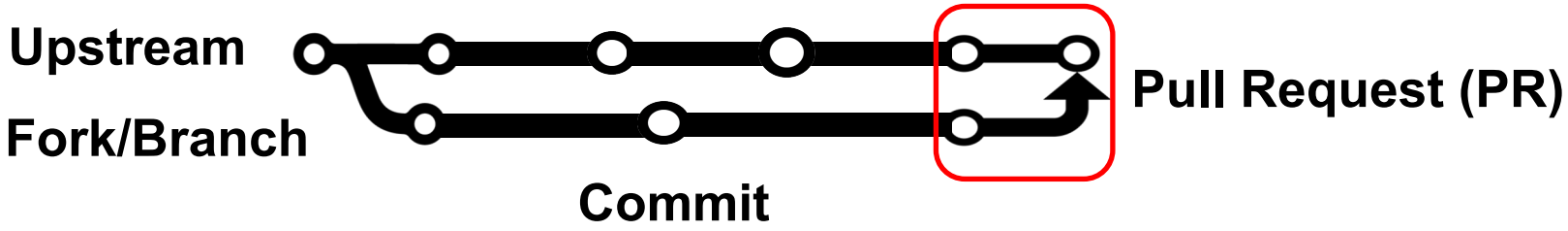
Fork-based Development

Fork-based Development

Fork-based Development



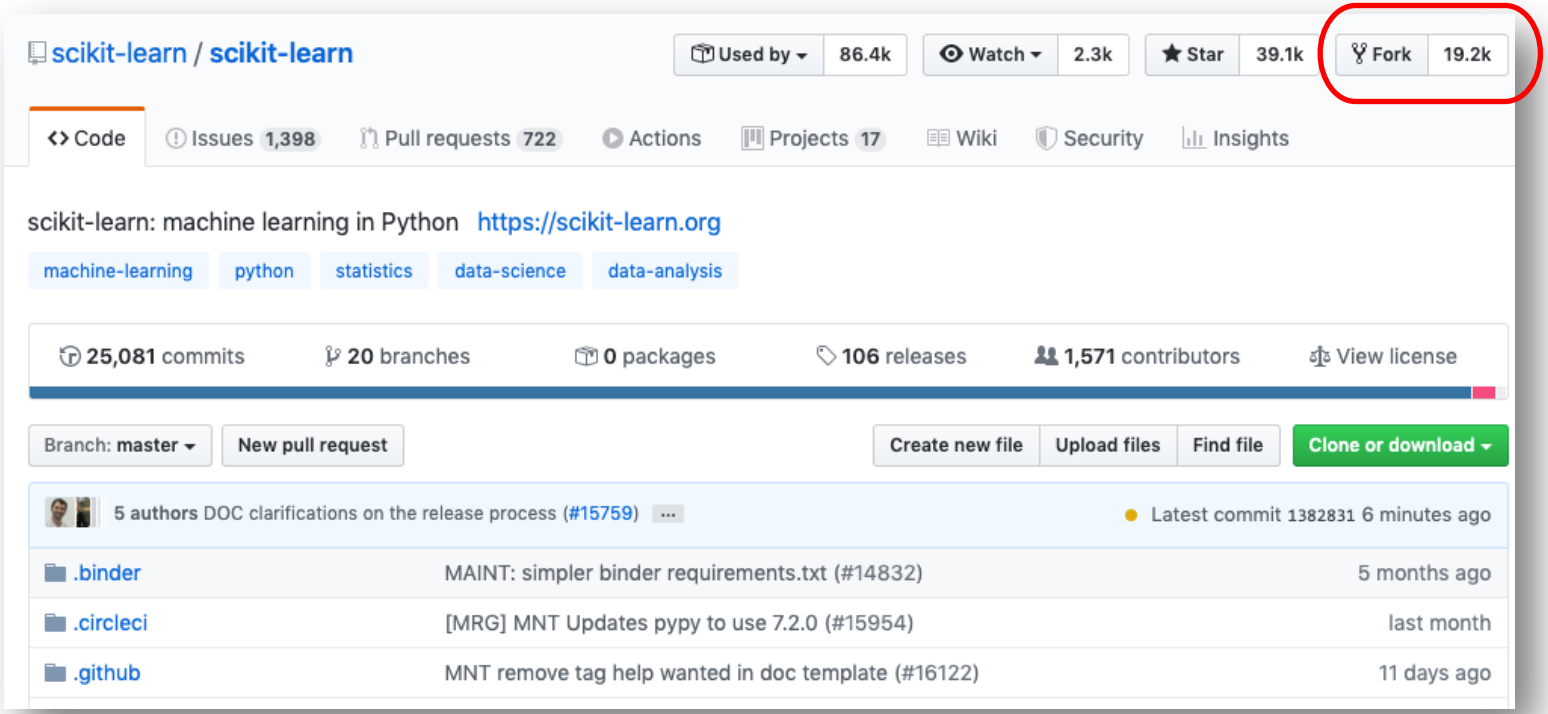
Fork-based Development



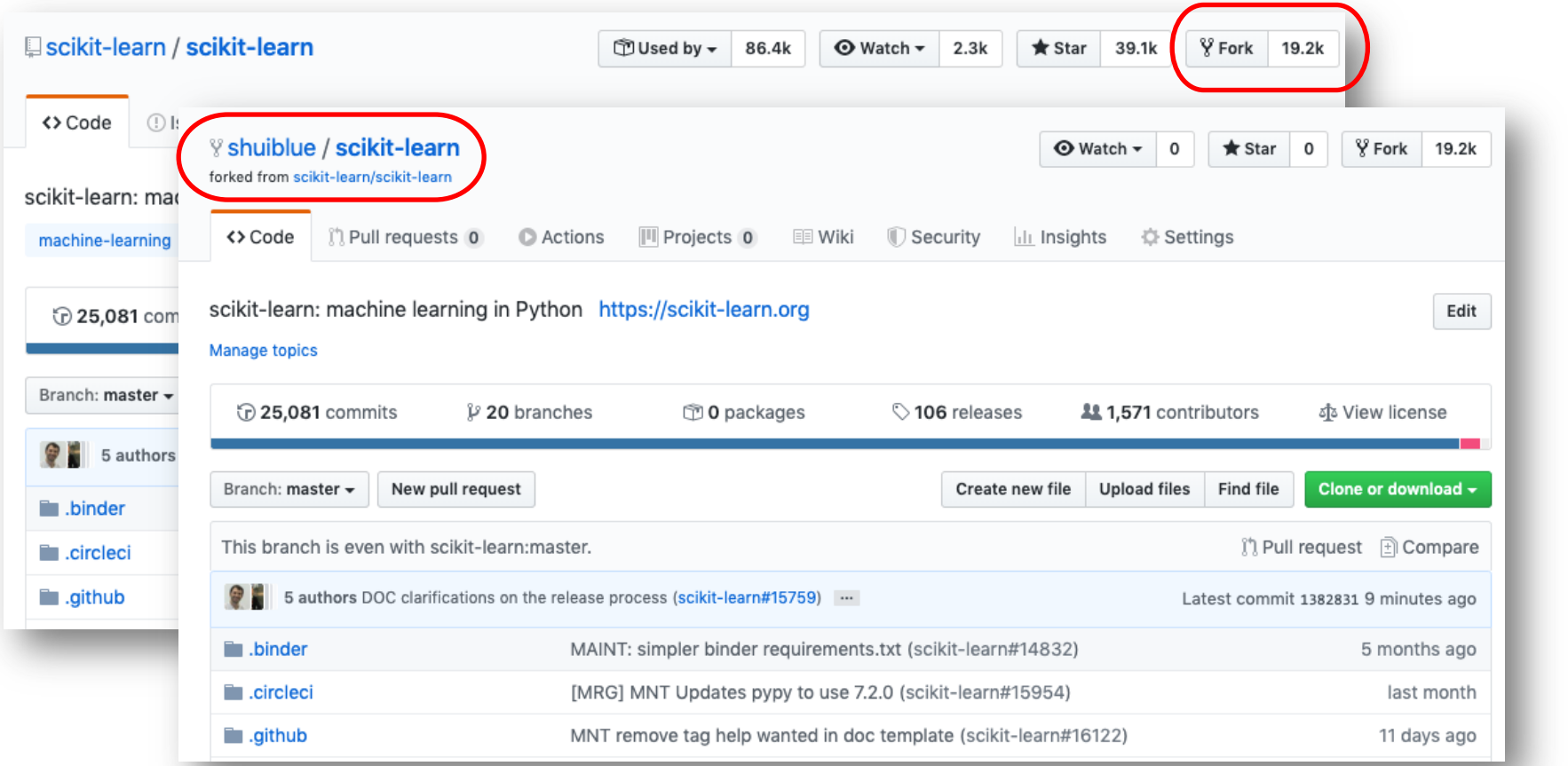
Fork-based / Branch-based / Pull-based Dev.

Pull Request / Merge Request

Fork-based Dev. Lowers Entry Barriers



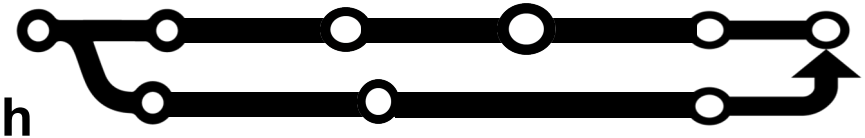
Fork-based Dev. Lowers Entry Barriers



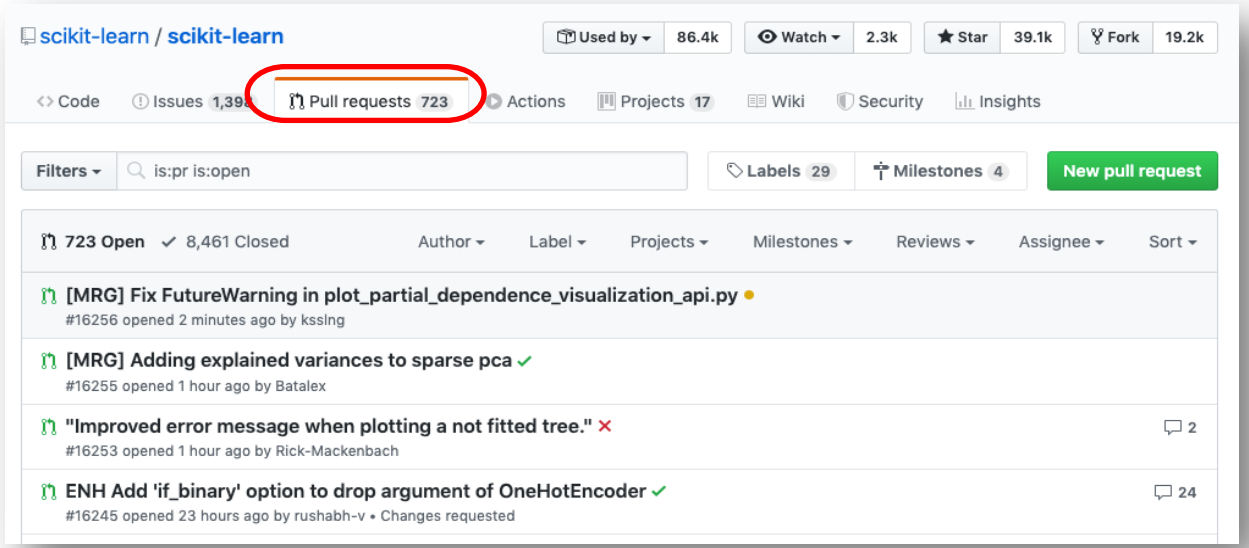
Fork-based Dev. Lowers Entry Barriers

Upstream

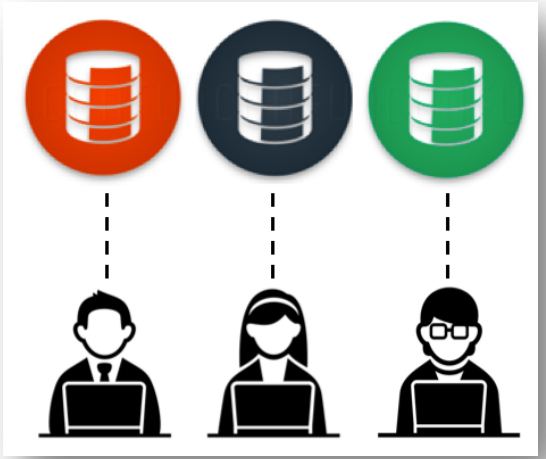
Fork/Branch



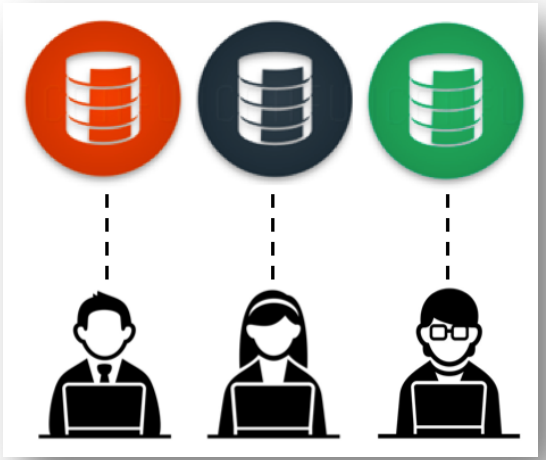
Pull Request (PR)



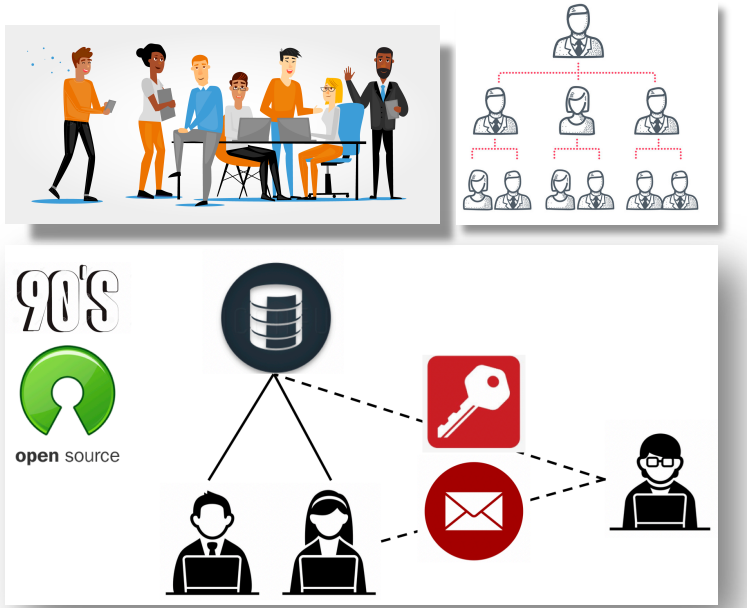
Fork-based Development



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]

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

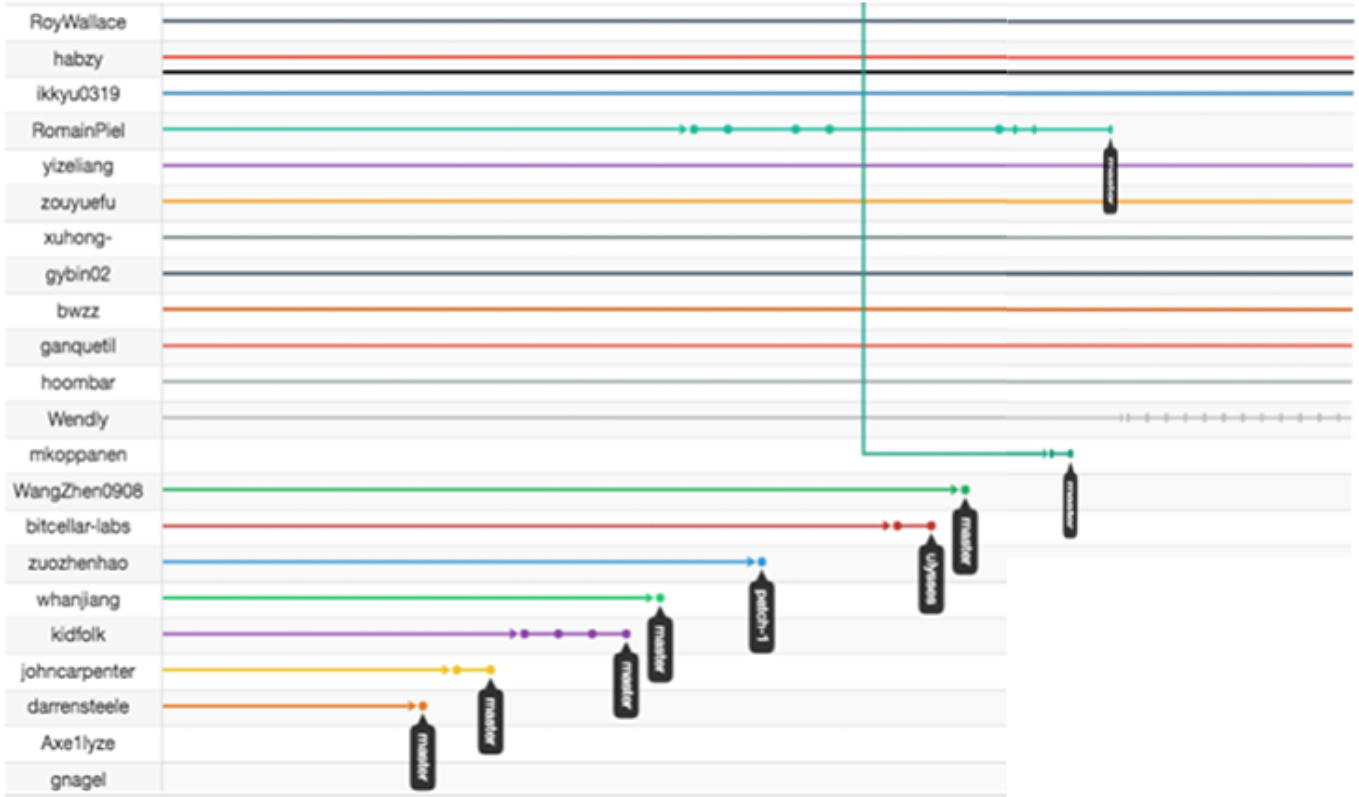
[GHTorrent 2019-06]

Fork-based Dev. Becomes Popular

https://github.com/customer-stories?type=enterprise

But

Problem -- Lost Contributions



Problem -- Redundant Development

foosel commented on Aug 22, 2017 — Owner

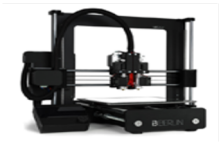
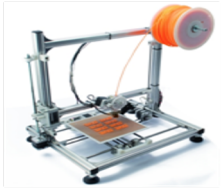
Sorry, but I can't stop laughing right now. I added *exactly* the same kind of functionality yesterday (just with a configurable ambient value and a debug command to also modify it during run time). See `fbcbb3f`

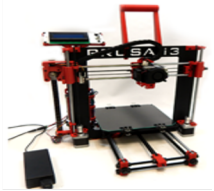
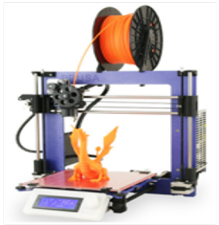
I can't believe this coincidence XD

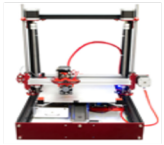
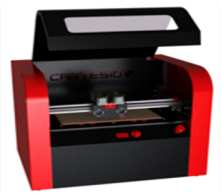
Noiredd commented on Nov 3, 2017 — Member

Duplicate of #5869 and #5972, partially also #5879.

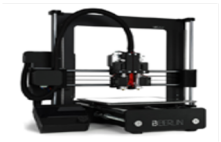
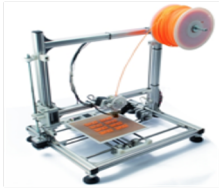
Problem -- Fragmented Community

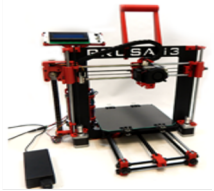
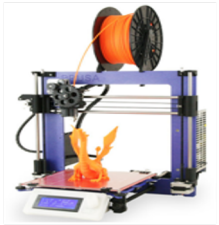



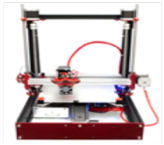
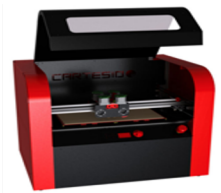



Problem -- Fragmented Community

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

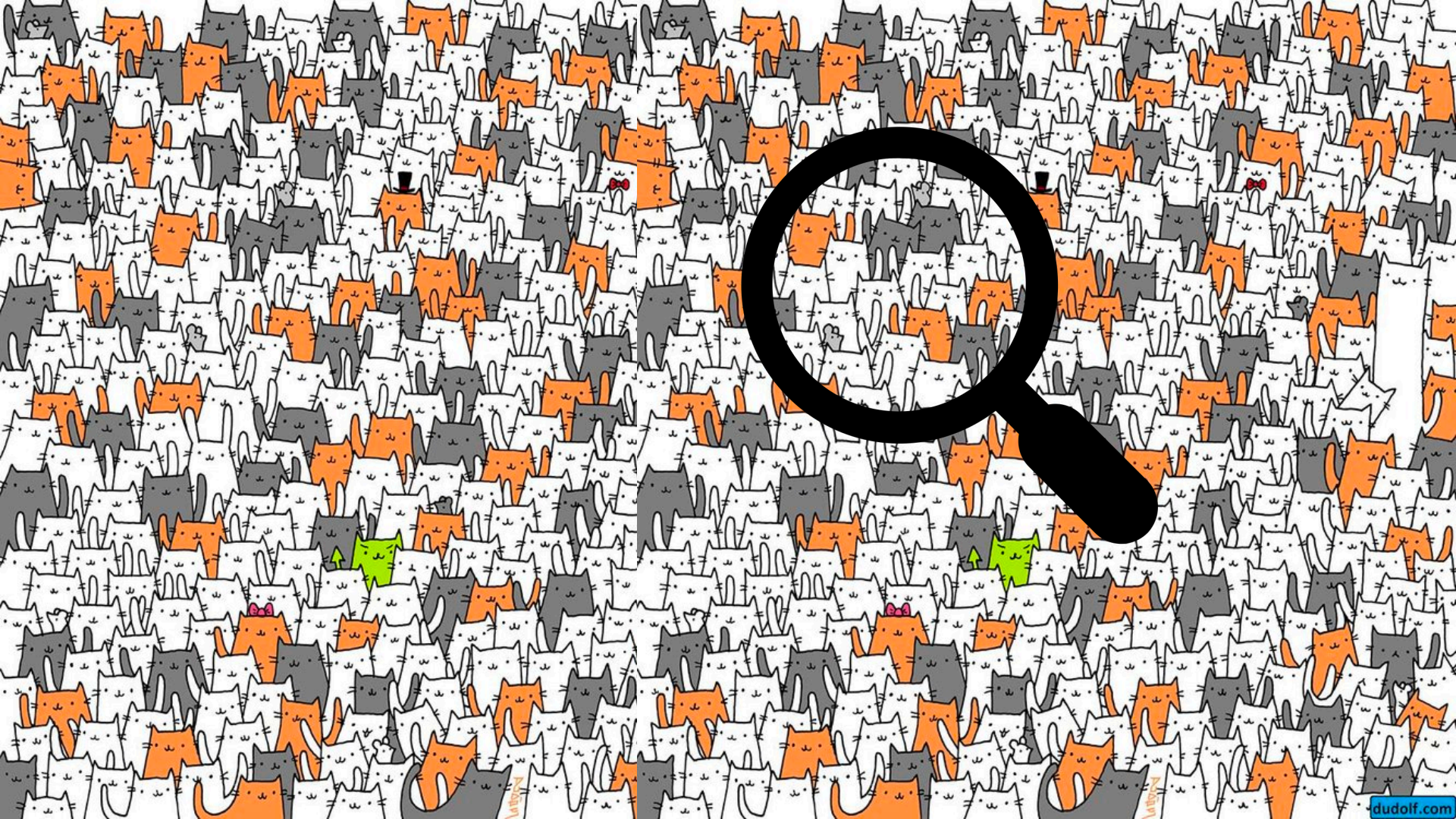
Fragmented Community

Is duplicate always bad?

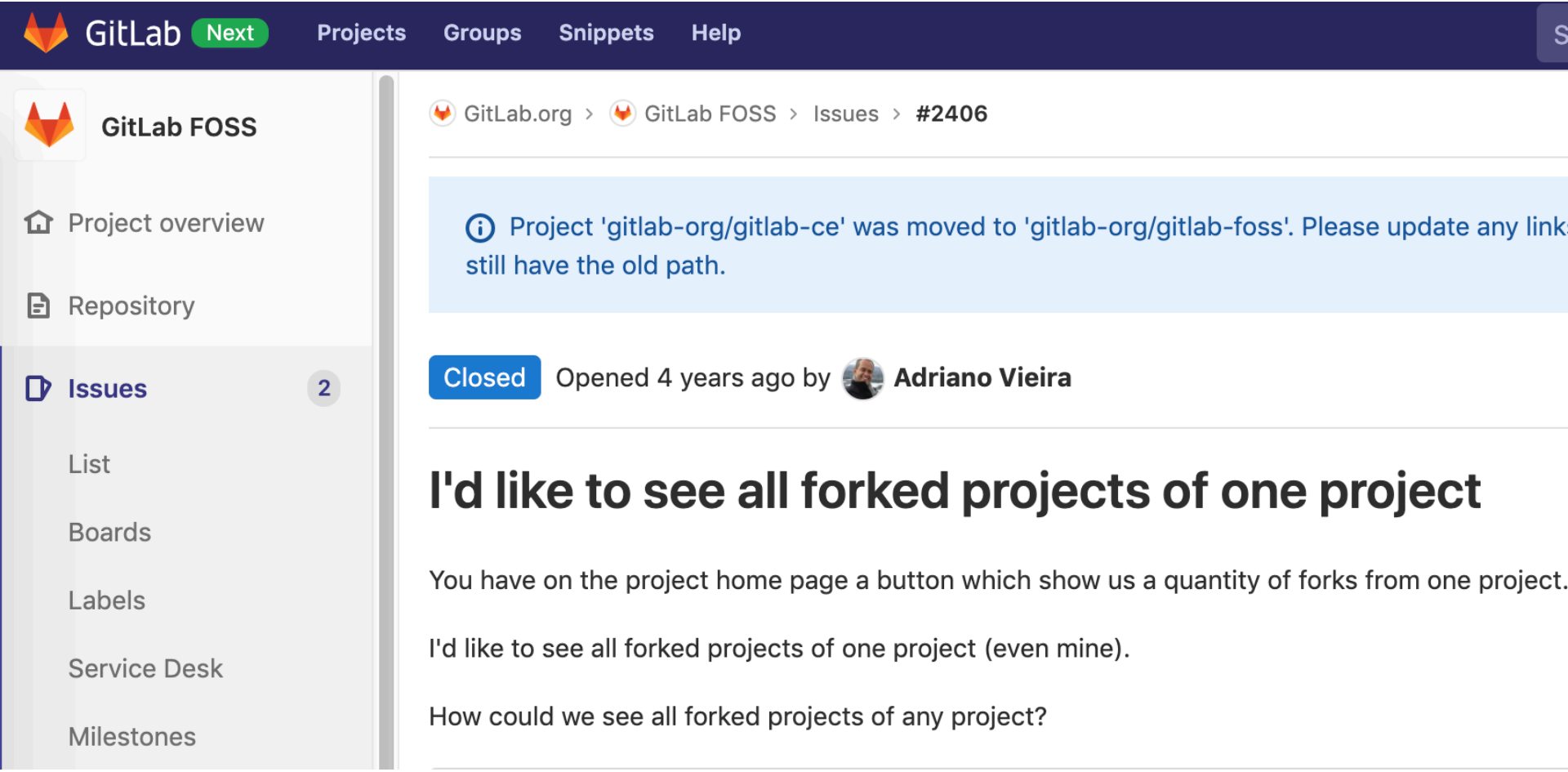
Similar Problems Happen in Industry

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

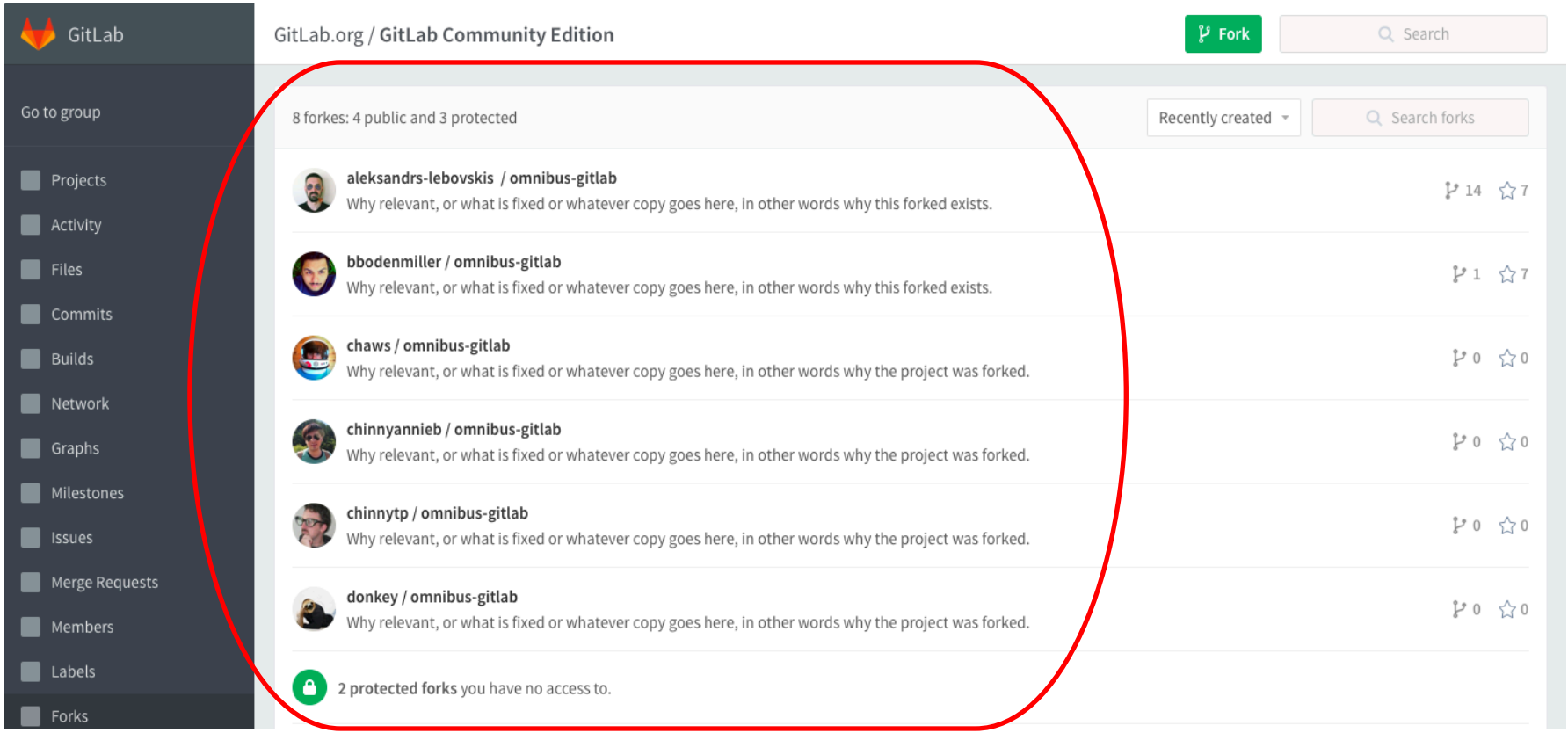

dudolf.com

Problem


GitLab
Next
Projects
Groups
Snippets
Help
GitLab FOSS
Project overview
Repository
Issues
2
List
Boards
Labels
Service Desk
Milestones
GitLab.org
GitLab FOSS
Issues
#2406
Project 'gitlab-org/gitlab-ce' was moved to 'gitlab-org/gitlab-foss'. Please update any link
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?

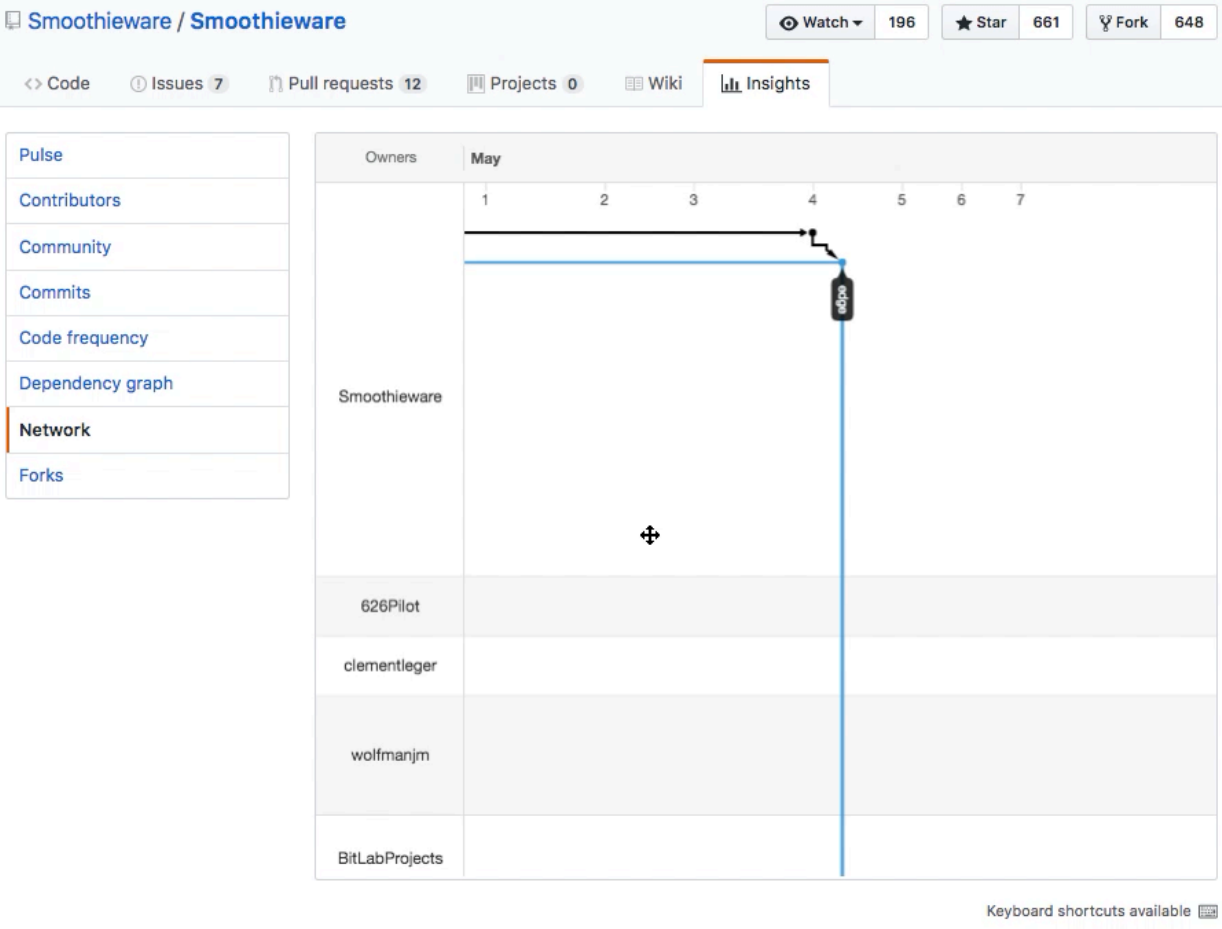
Problem

List of Forks



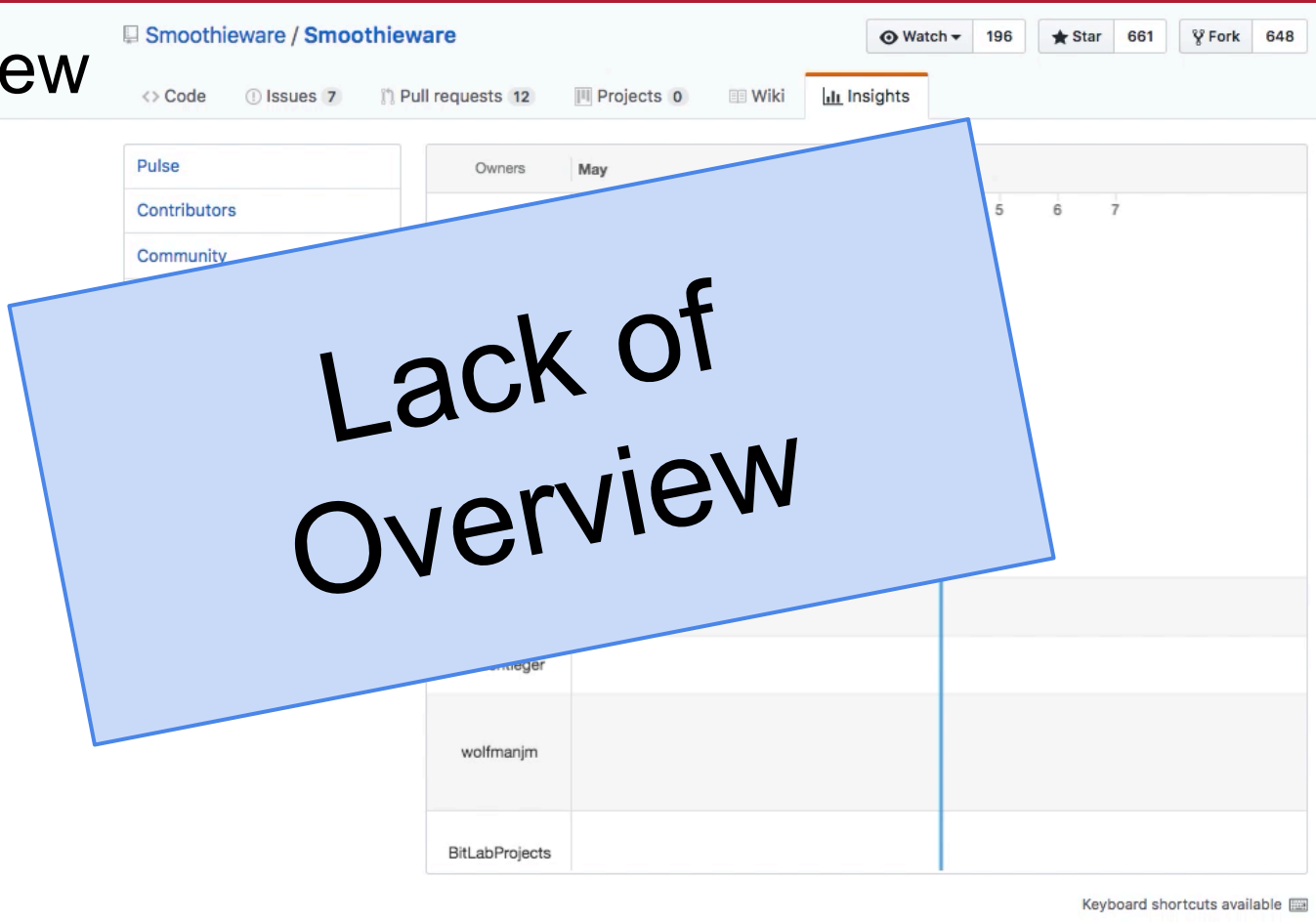
Problem

Network View

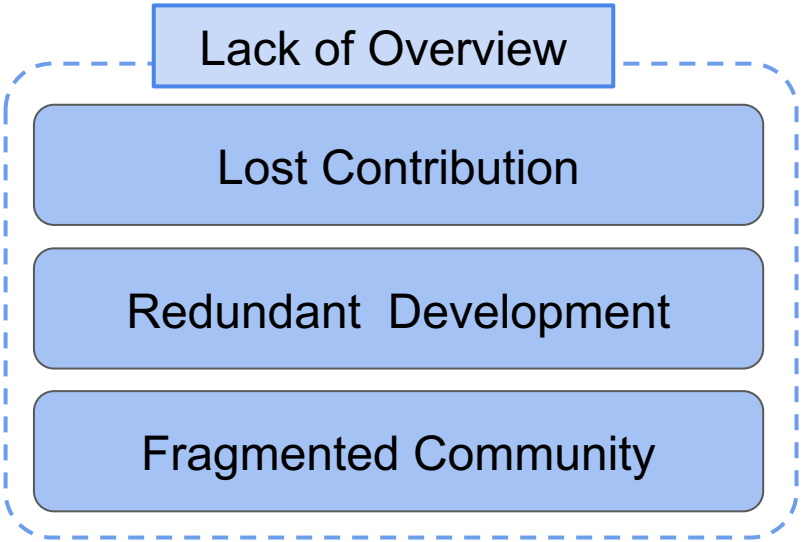


Problem

Network View



Problem


Lack of Overview
Lost Contribution
Redundant Development
Fragmented Community

Improving Collaboration Efficiency

Software Dev.

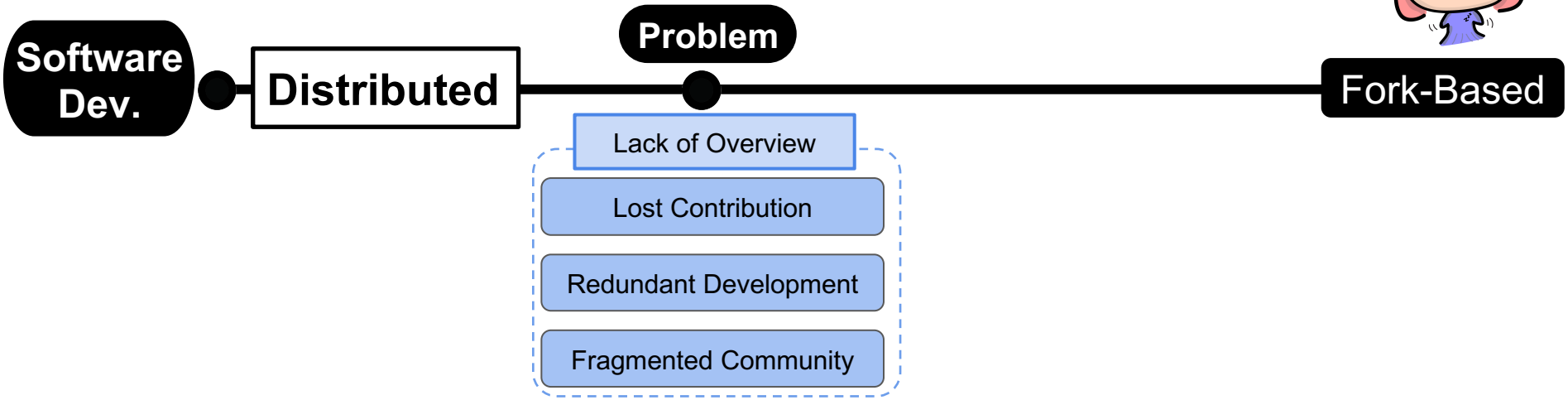
Improving Collaboration Efficiency

Software Dev. — **Distributed**

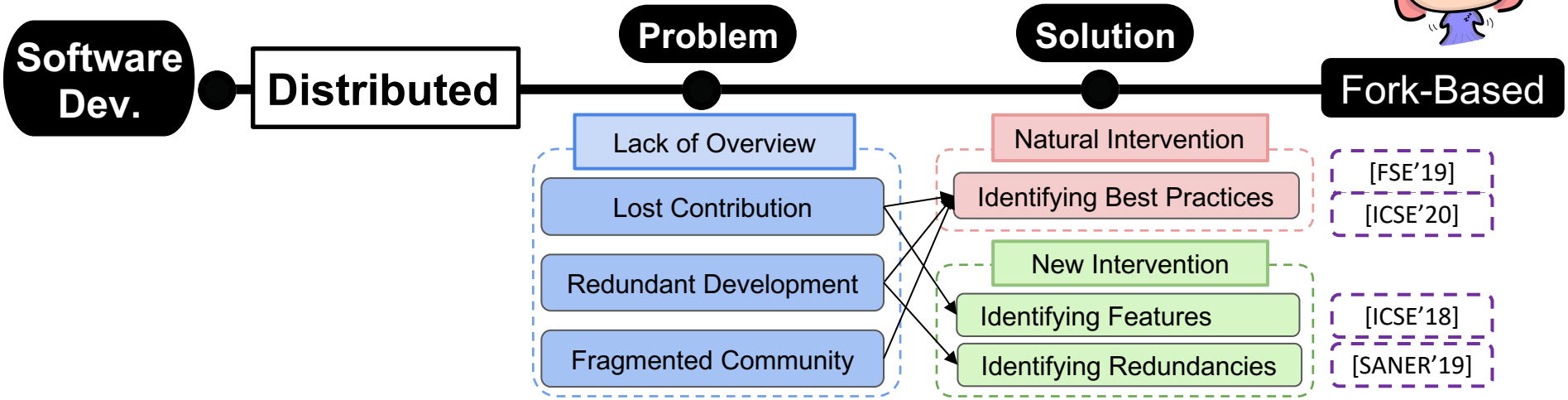
Improving Collaboration Efficiency

Software Dev. — Distributed — Fork-Based

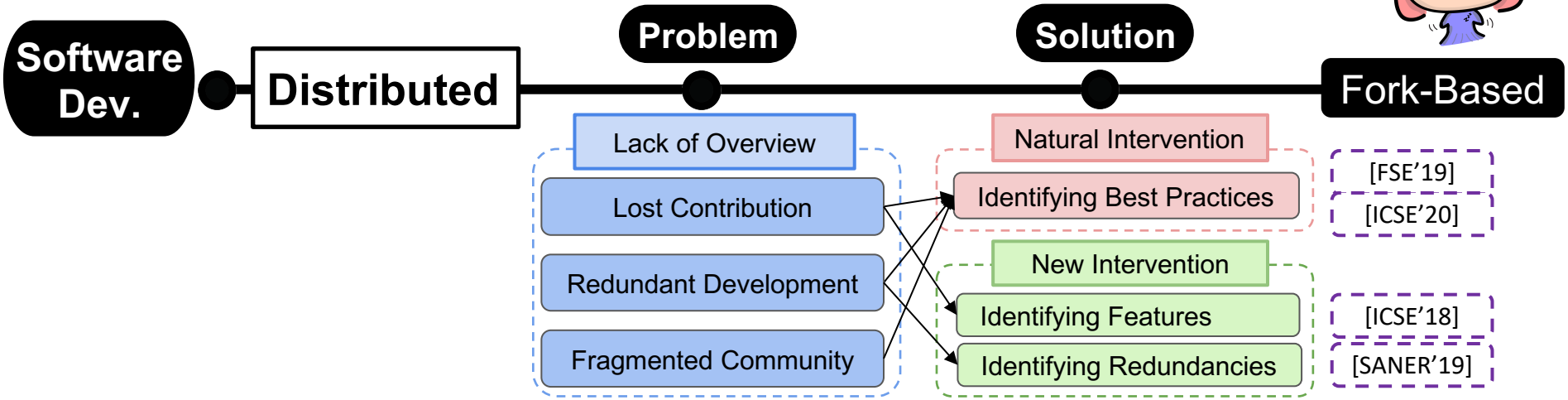
Improving Collaboration Efficiency



Improving Collaboration Efficiency



Improving Collaboration Efficiency



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

[FSE'19]

[ICSE'20]

New Intervention

Identifying Features

[ICSE'18]

Identifying Redundancies

[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

[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]

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

- Project proposal
- Resolve issues on the issue tracker

- Open for any contribution

Projects are different

- 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

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

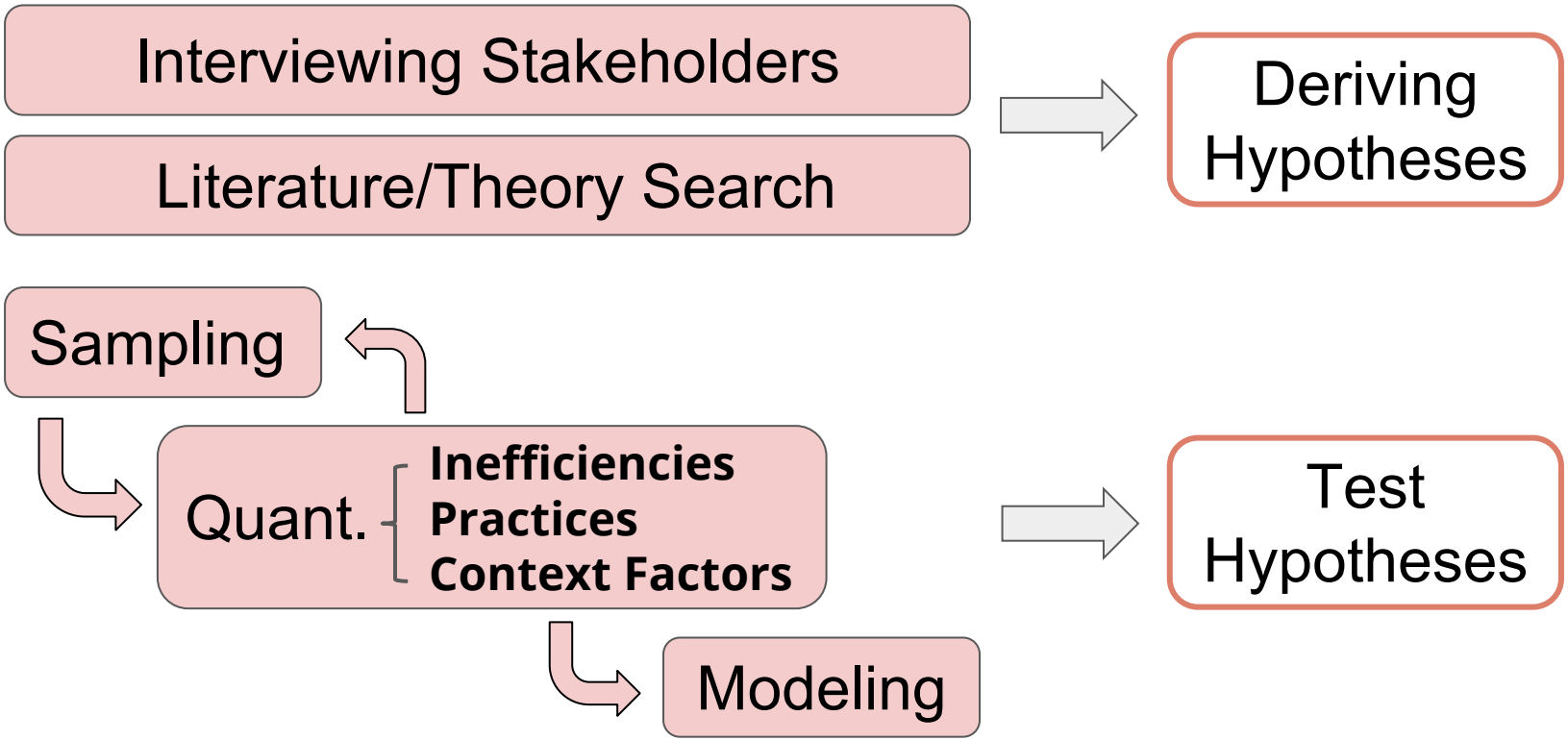
Cross-sectional Correlational Study

- 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

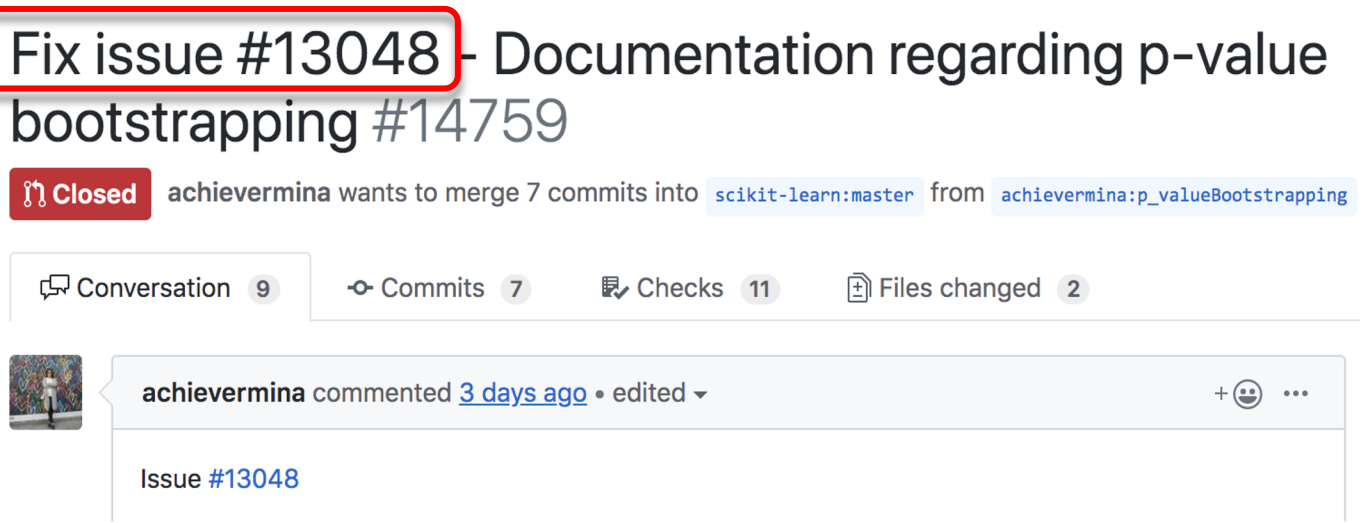
Quantifying
- Inefficiencies
- Practices
- Context Factors

Hypo: Centralized Management ➔ Larger portion of contributing forks

Regression Modeling

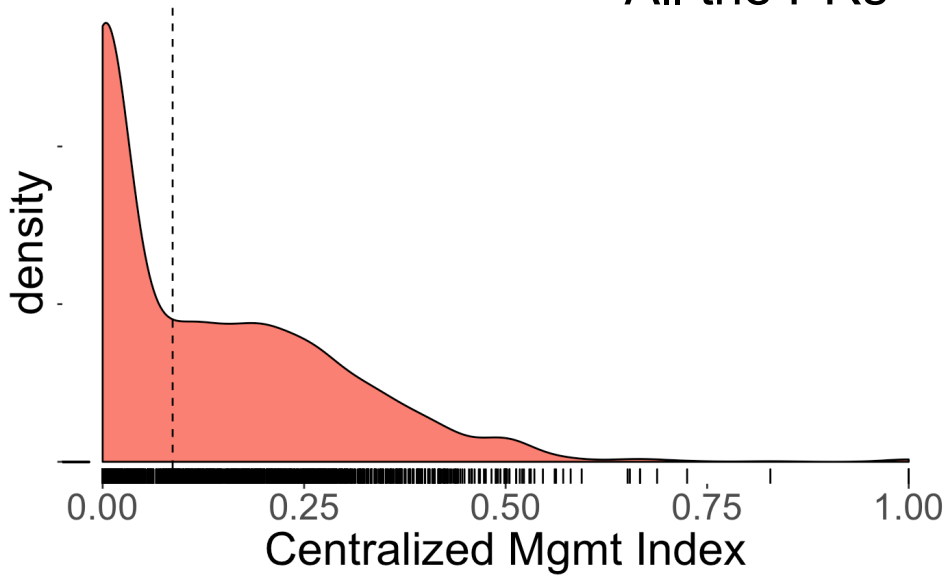
Operationalization - Centralized Management

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



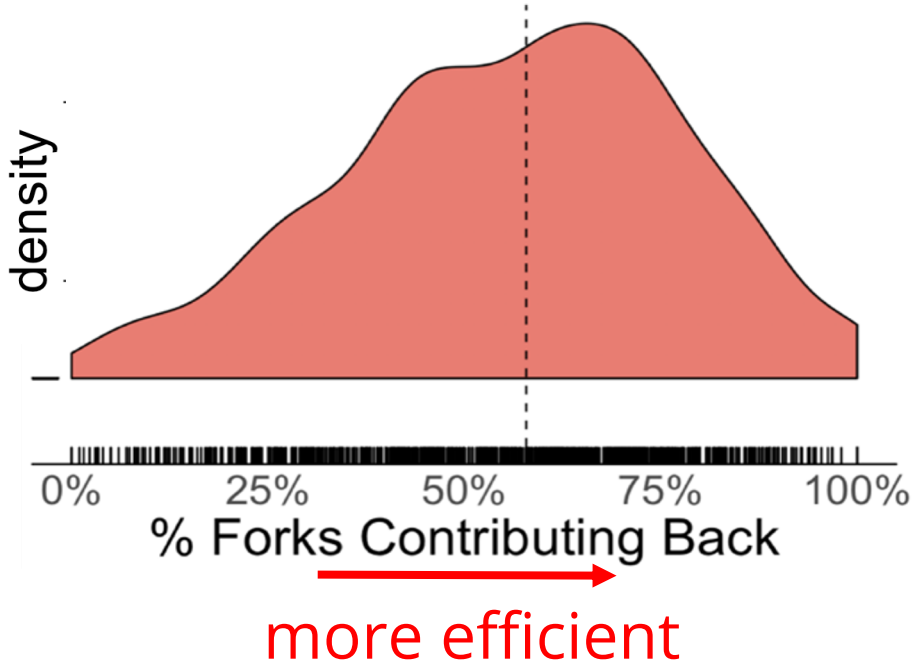
Operationalization - Centralized Management

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

Sampling

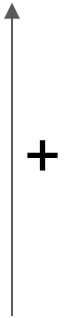
Group	#fork	#projects	#projects in sample set
A	[3,000 , +]	231	200
B	[1,000 , 3,000)	847	300
C	[20 , 1,000)	116,532	1300

Quantifying
- Inefficiencies
- Practices
- Context Factors

Multiple Regression Modeling

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.

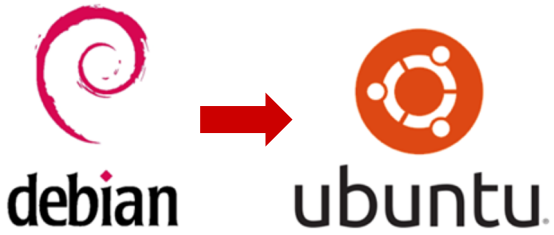
➔

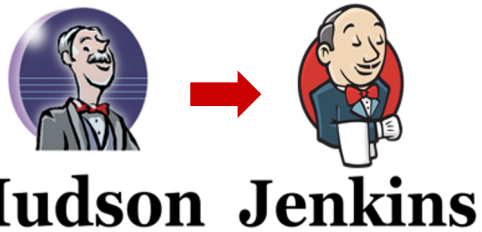
Old Notion of Forking: Splitting off a Community

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.



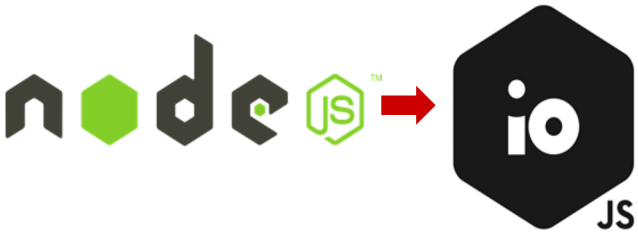


Community Fragm. is Expensive

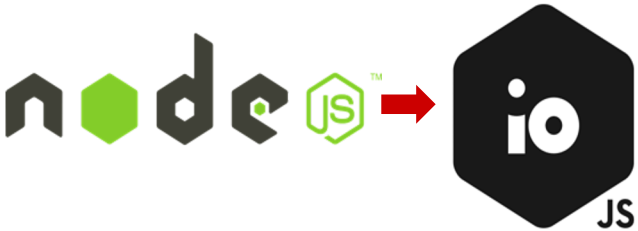
Forking was a Weighty Decision

A strong norm against forking [Yoo 2016]

Community Fragm. is Expensive

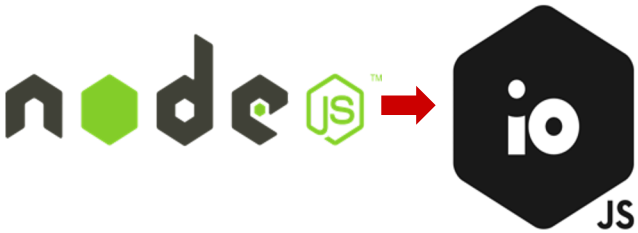


Community Fragm. is Expensive



"Some open-source forks have made life difficult for developers. … that will force developers to pick sides." -Lauren Orsini

Community Fragm. is Expensive

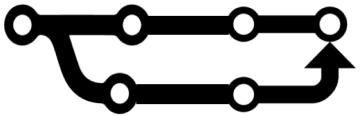


"Some open-source forks have made life difficult for developers. … that will force developers to pick sides." -Lauren Orsini

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

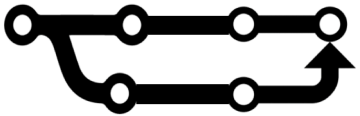
by OWEN WILLIAMS — Jun 16, 2015 in DESIGN & DEV

Different kinds of Forks



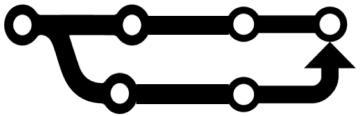
Hard Fork VS **(Social) Fork**

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

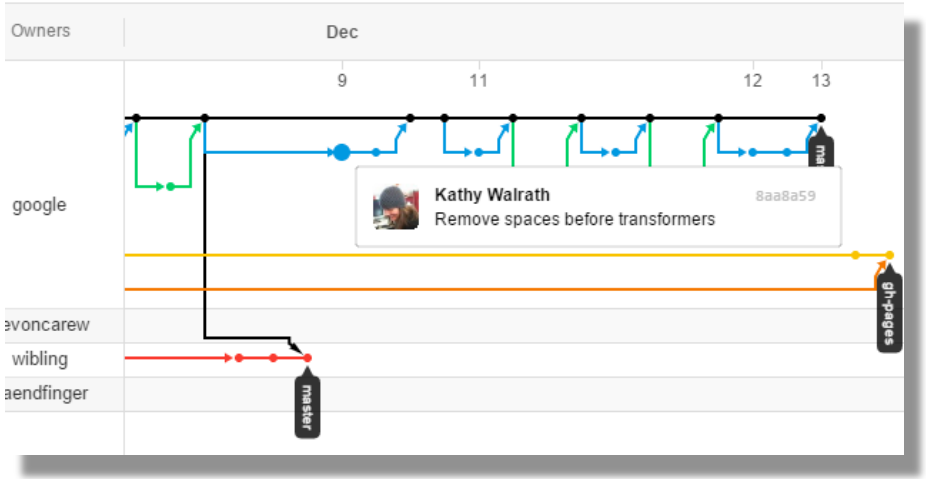
Shurui Zhou
Carnegie Mellon University, USA

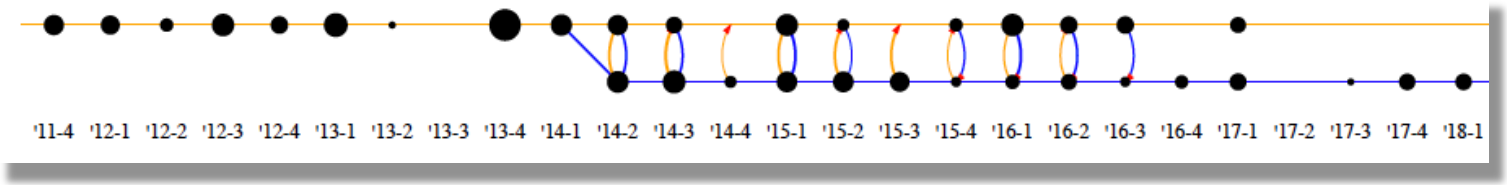
Bogdan Vasilescu
Carnegie Mellon University, USA

Christian Kästner
Carnegie Mellon University, USA

Operationalization – Community Fragmentation

Detecting Hard Forks



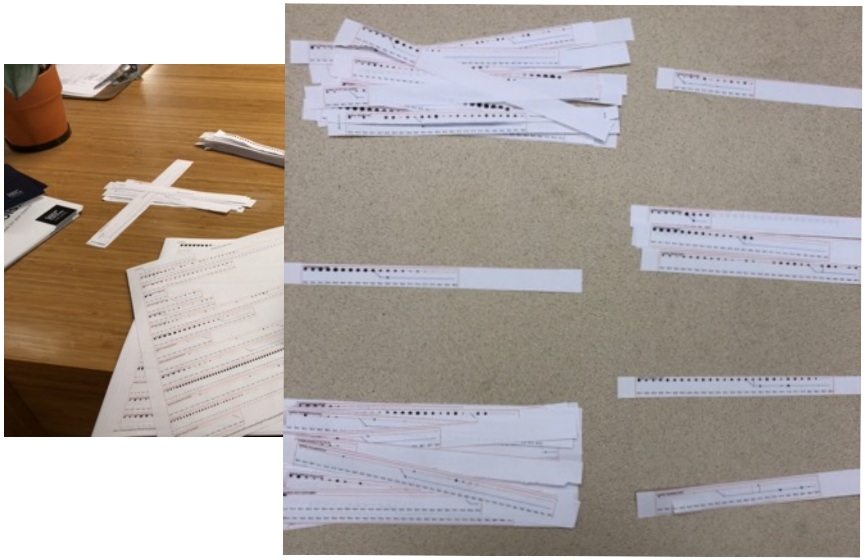


Operationalization – Community Fragmentation

Identifying Evolution Patterns of Hard Forks

Operationalization – Community Fragmentation

Identifying Evolution Patterns of Hard Forks



Operationalization – Community Fragmentation

Identifying Evolution Patterns of Hard Forks

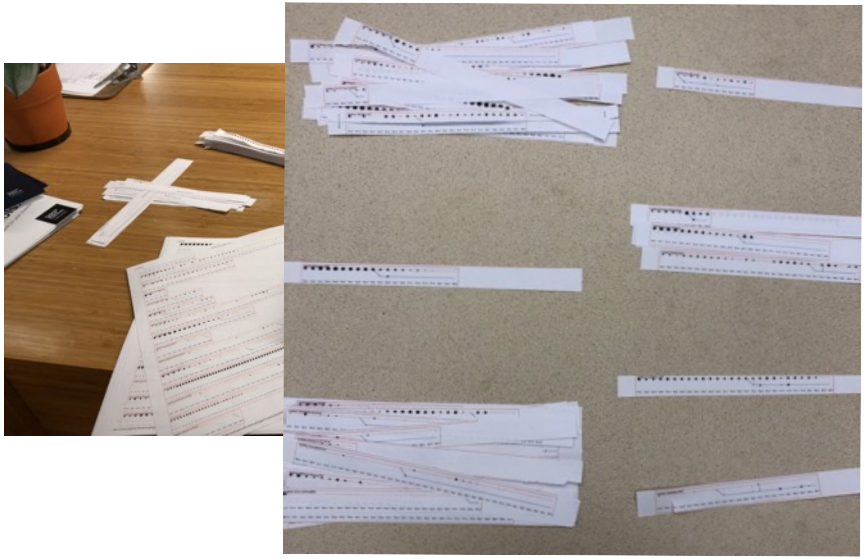


Table 2: Evolution patterns of hard forks

Id		Category	Total	Sub-category	Example	Count	Interviewees
1	Revive Dead Project	Success (F. active > 2 Qt.)	632	Upstream remains inactive		576	P12
2				Upstream active again		56	
3		Not success (F active <= 2 Qt)	420			420	
4	Forking Active Project	Both Alive	723	only merge		26	P10
5				only sync		107	P2, P13, P15
6				merge & sync		28	P9
7				no interation		562	P1, P3, P4, P5, P7, P14
8		Fork Lived Longer	7280	only merge		174	
9				only sync		686	
10				merge & sync		107	
11				no interaction		6313	P6, P8, P11
12		Fork does not out live upstream	6251	only merge		388	
13				only sync		762	
14				merge & sync		199	
15				no interaction		4902	

Operationalization – Community Fragmentation

Identifying Evolution Patterns of Hard Forks

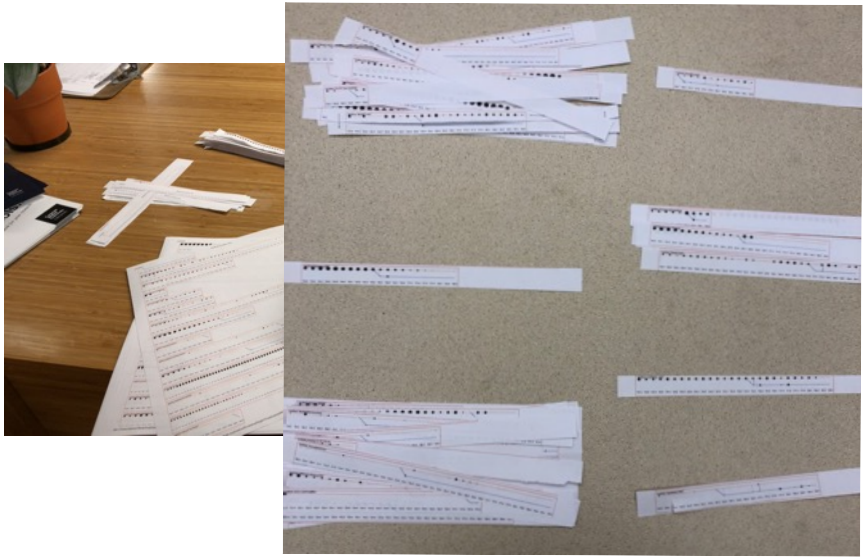
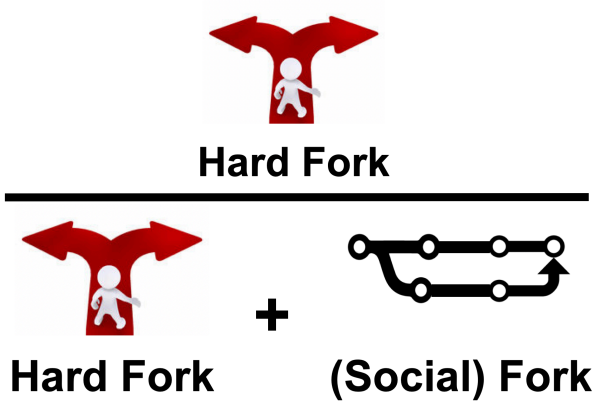


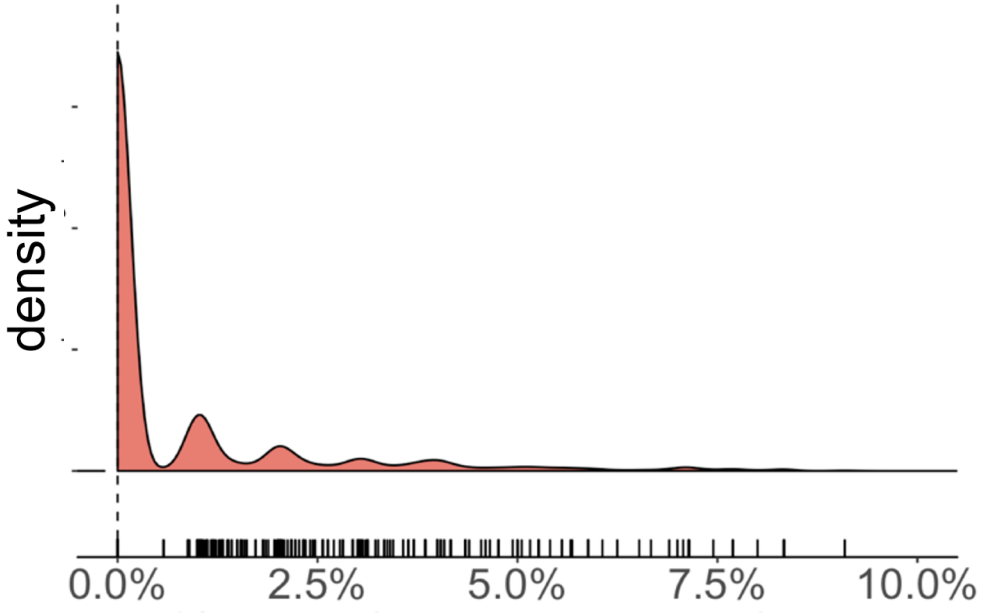
Table 2: Evolution patterns of hard forks

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

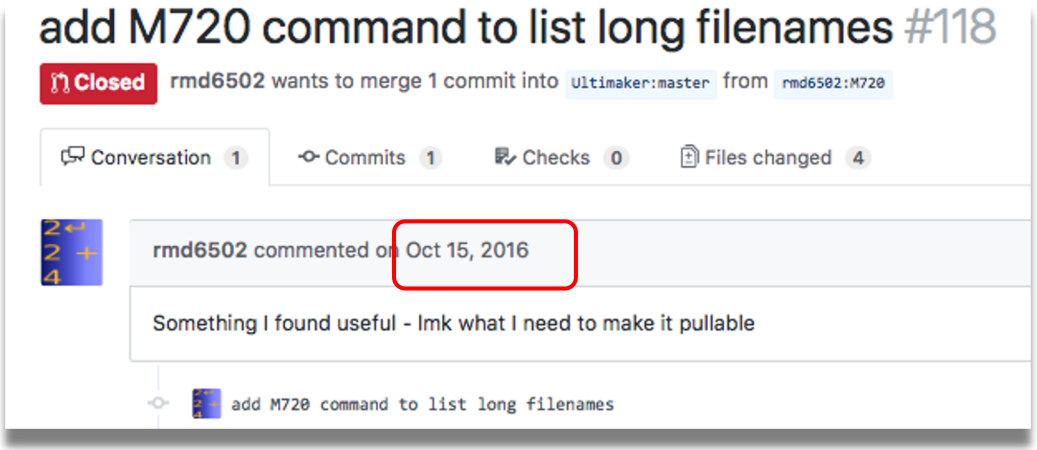
Ultimaker

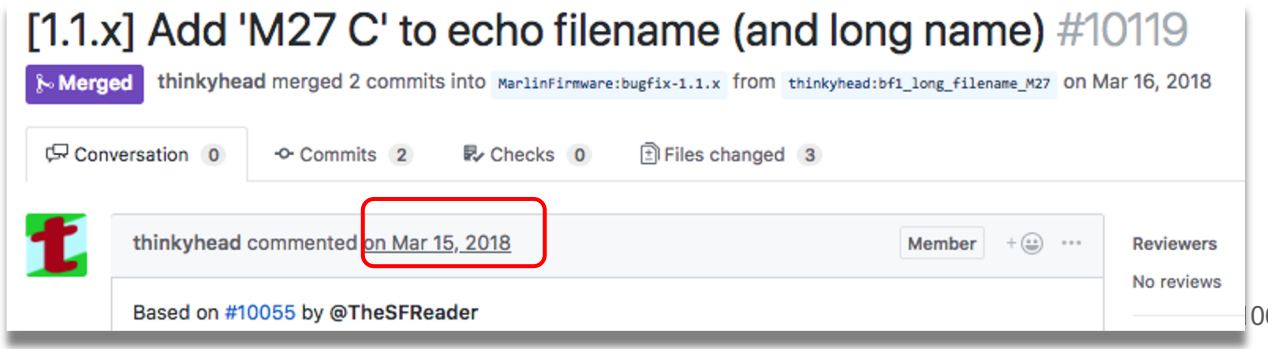
Ultimaker / Ultimaker2Marlin

Hard Fork

Marlin

MarlinFirmware / Marlin





Hypoth: Centralized Mgmt → Community Fragm.

Community Fragmentation

\+

Centralized Management

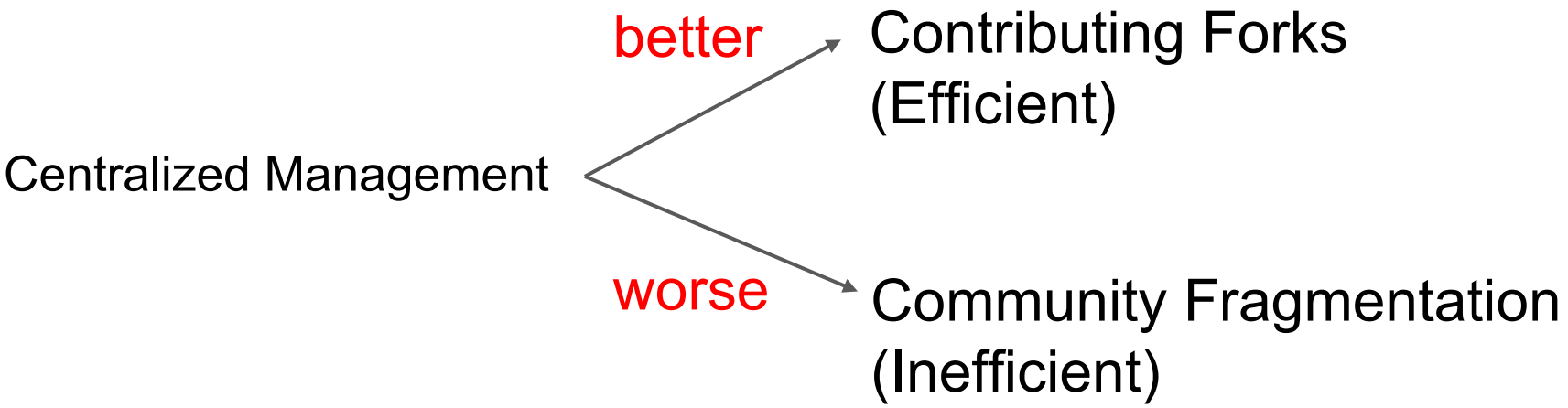
(12% of variance explained)

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

Problem

Solution

Lack of Overview

- Lost Contribution
- Redundant Development
- Fragmented Community

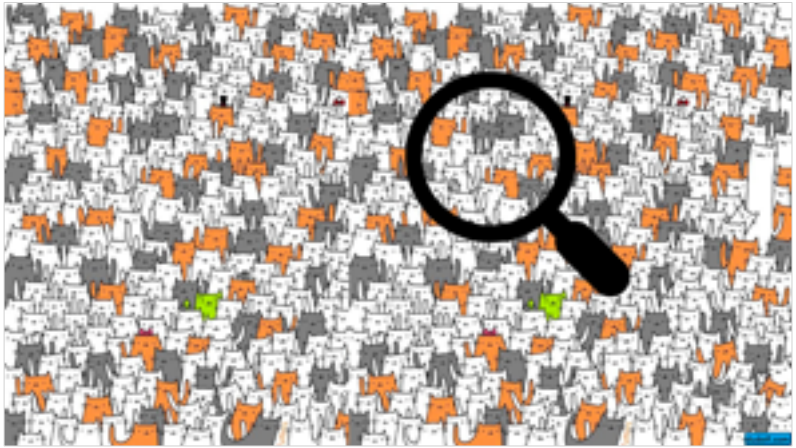
Natural Intervention

- Identifying Best Practices — [FSE'19] [ICSE'20]

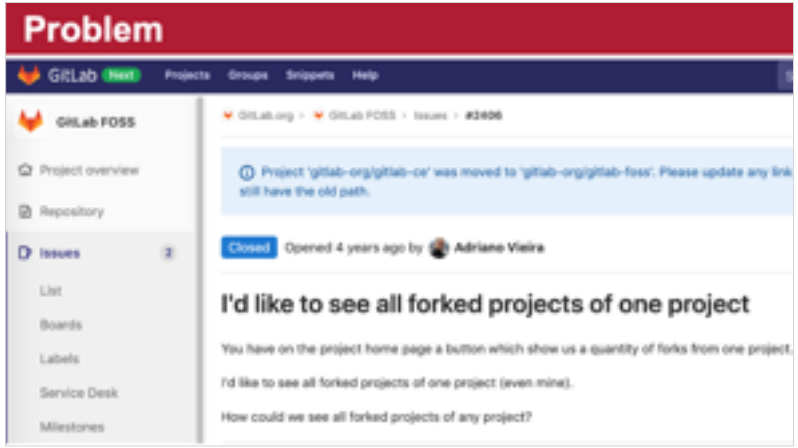
New Intervention

- Identifying Feature — [ICSE'18]
- Identifying Redundancies — [SANER'19]

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

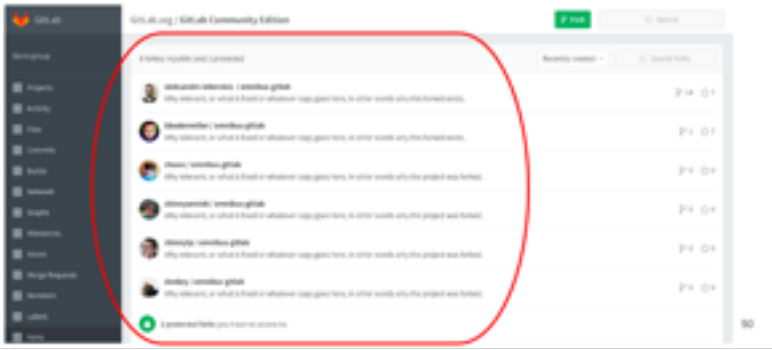


Problem



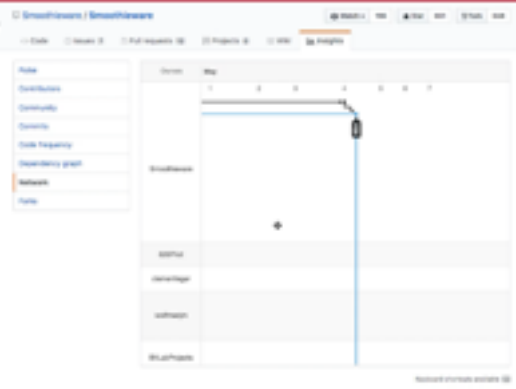
Problem

List of Forks



Problem

Network View



Problem

Solution

Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Natural Intervention

Identifying Best Practices

[FSE'19]

[ICSE'20]

New Intervention

[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

[FSE'19]

[ICSE'20]

New Intervention

Awareness Tools

[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

Ştefan 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?

INFOX

Summarizing forks that has un-merged commits

Mapping between feature to code changes

INFOX

Email System

- Signature
- Encryption

Feature id	*Keyword List*	*LOC*
Sig.	Signature, isSigned, ...	23
Enc.	Encryption, Decryption, isEncrypted, Decrypt, ...	100

INFOX

Email System

- Signature
- Encryption
- Decryption

Feature id	*Keyword List*	*LOC*
Sig.	Signature, isSigned, ...	23
Enc.	Encryption, isEncryped, ...	55
Dec	Decryption, Decrypt, ...	45

INFOX

Feature	Navigation		Keyword List	LOC
Sig.	Prev.	Next	Signature, isSigned, ...	23
Enc.	Prev.	Next	Encryption, Decryption, isEncryped, 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

Feature	ibradypod/phantomjs, last commmit:May 28	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	raff/phantomjs, last commmit: Mar 5	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	ricokahler/phantomjs, last commmit:Feb 2	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	DeviaVir/phantomjs, last commmit: Jan 25, 2016	LOC
allow	allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio	7

INFOX Overview Page

Feature	ibradypod/phantomjs, last commmit:May 28	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

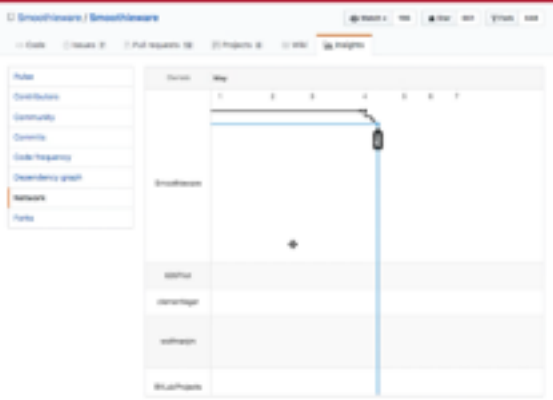
Feature	raff/phantomjs, last commmit: Mar 5	LOC
dow.	download, com, pull, file, ad, support, ariya, http	39
get	get, qt, are, kei_enter, el, mouse, require, clicks, hard, absolute, setfocus, button, coordinates, keypress	29

Feature	ricokahler/phantomjs, last commmit:Feb 2	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

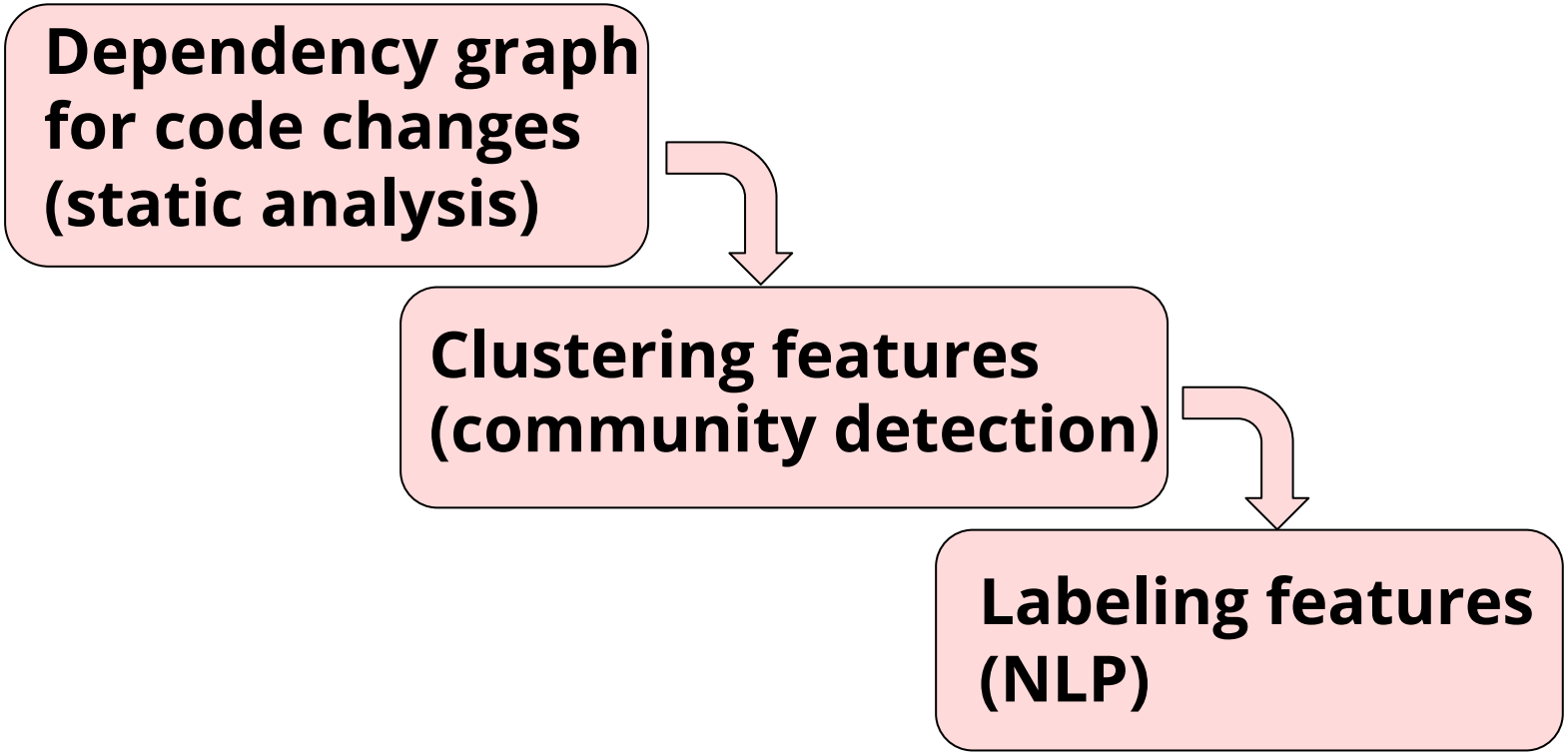
Feature	DeviaVir/phantomjs, last commmit: Jan 25, 2016	LOC
allow	allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio	7

Problem

Network View



INFOX


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);
```

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);
```

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);
```

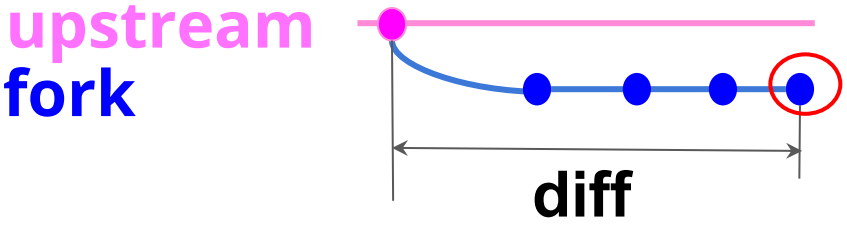
File 2: Email.c

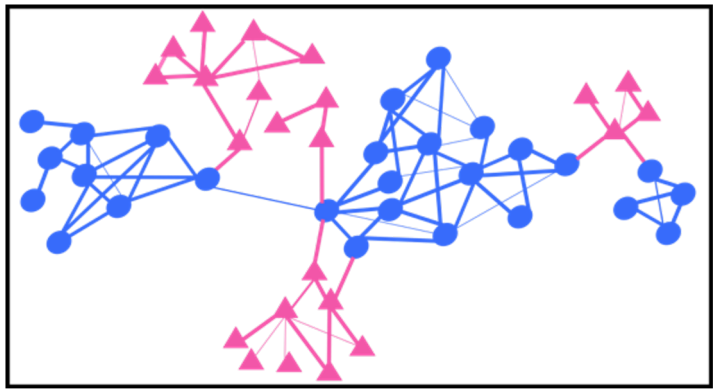
```
1  +  void printMail ( struct email *msg)
2     {
3         printf ("SUBJECT:", msg -> subject );
4  +      printf ("SIGNED:", msg->isSigned);
5  +      if (0 == (isEncrypted(msg) ) )
6             printf ( "BODY:", msg -> body );
7  +      else
8  +          printf ( "Encrypted msg." );
9     }
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 +  }
```

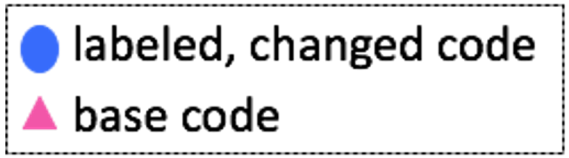
3 Dependencies

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

Dependency Graph





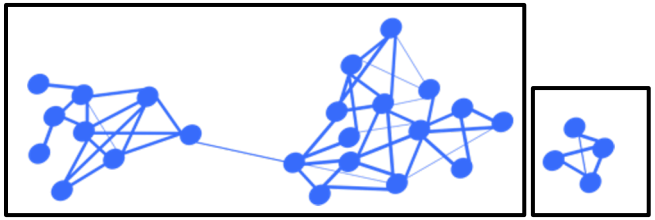


Dependency graph

Clustering features

Labeling features

Dependency Graph

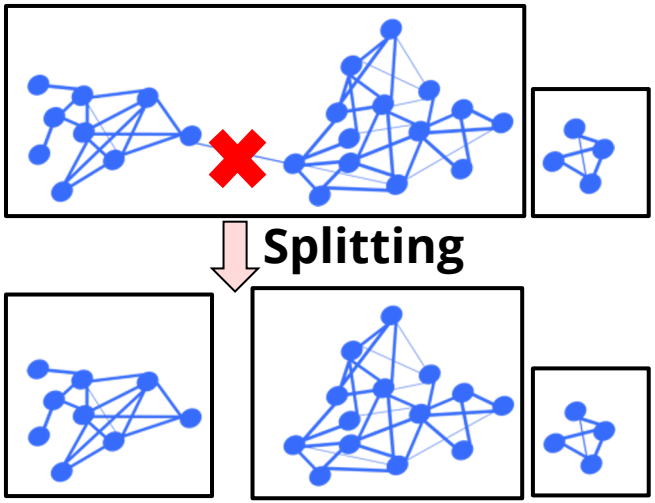


Dependency graph

Clustering features

Labeling features

Dependency Graph

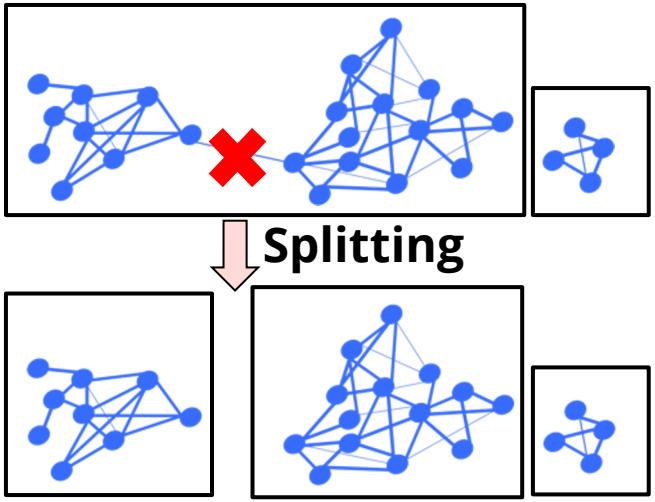


Dependency graph

Clustering features

Labeling features

Dependency Graph



Network Analysis
Girvan–Newman algorithm

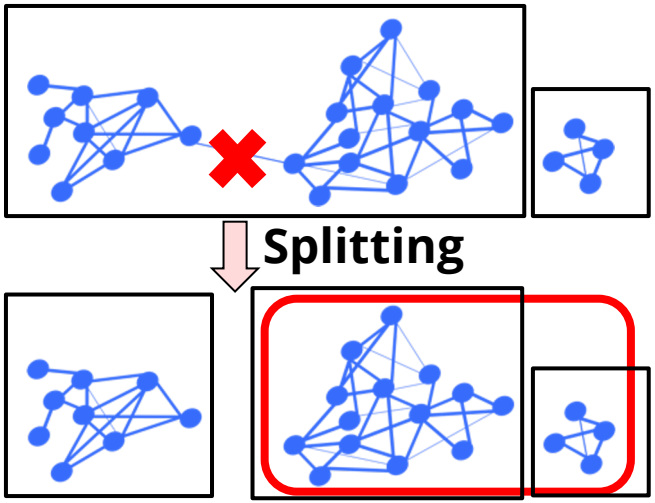
+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods

Dependency graph

Clustering features

Labeling features

Dependency Graph

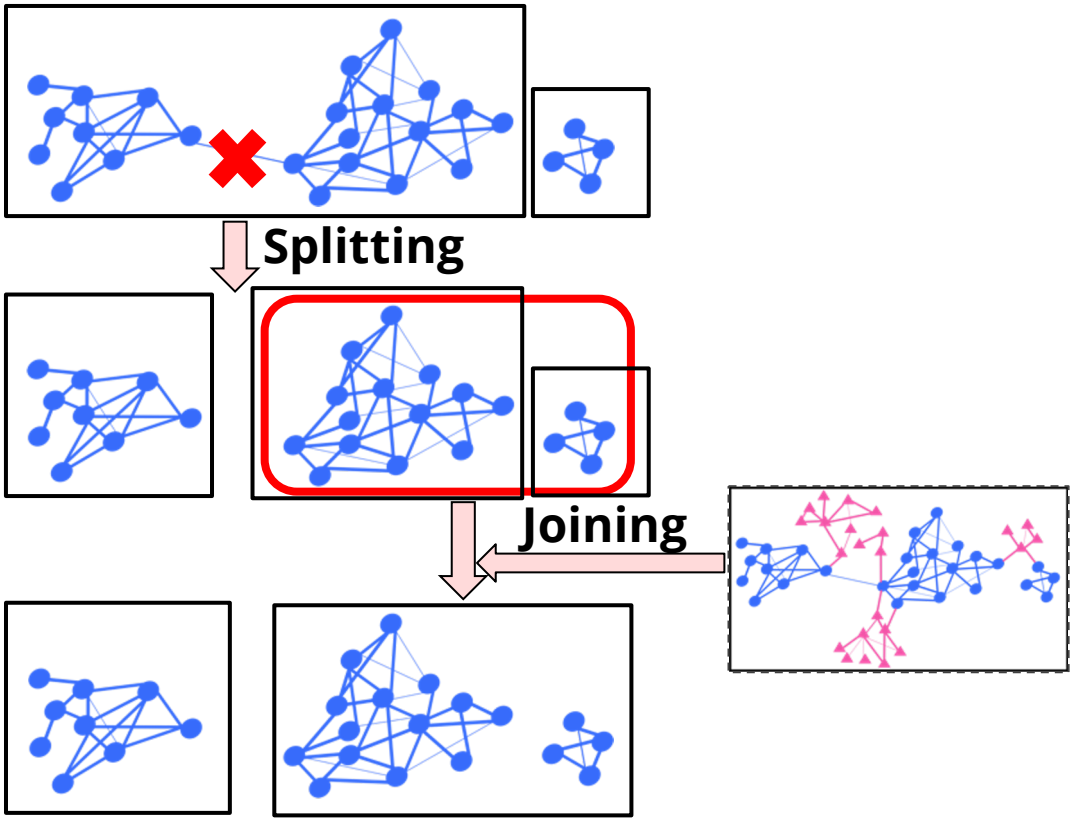


Dependency graph

Clustering features

Labeling features

Dependency Graph

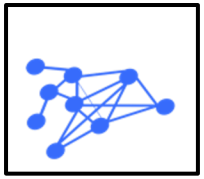


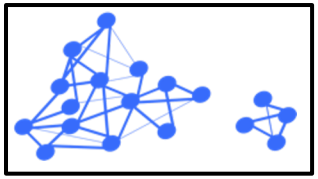
Dependency graph

Clustering features

Labeling features

Dependency Graph





- commit message
- code
- comment

↓ TF-IDF, N-Gram

Sig.	Signature, isSigned, ...	23
Enc.	Encryption, Decryption, isEncrpyed, Decrypt, ...	100

Dependency graph

↓

Clustering features

↓

Labeling features

INFOX

Feature	ibradypod/phantomjs, last commmit:May 28	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	raff/phantomjs, last commmit: Mar 5	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	ricokahler/phantomjs, last commmit:Feb 2	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	DeviaVir/phantomjs, last commmit: Jan 25, 2016	LOC
allow	allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio	7

INFOX

Feature	Navigation		Keyword List	LOC
Sig.	Prev.	Next	Signature, isSigned, ...	23
Enc.	Prev.	Next	Encryption, Decryption, isEncryped, 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 - Evaluation

Effectiveness

Usefulness

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

INFOX - Effectiveness

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

INFOX - Effectiveness

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

Quantitative Study

INFOX - Effectiveness

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	

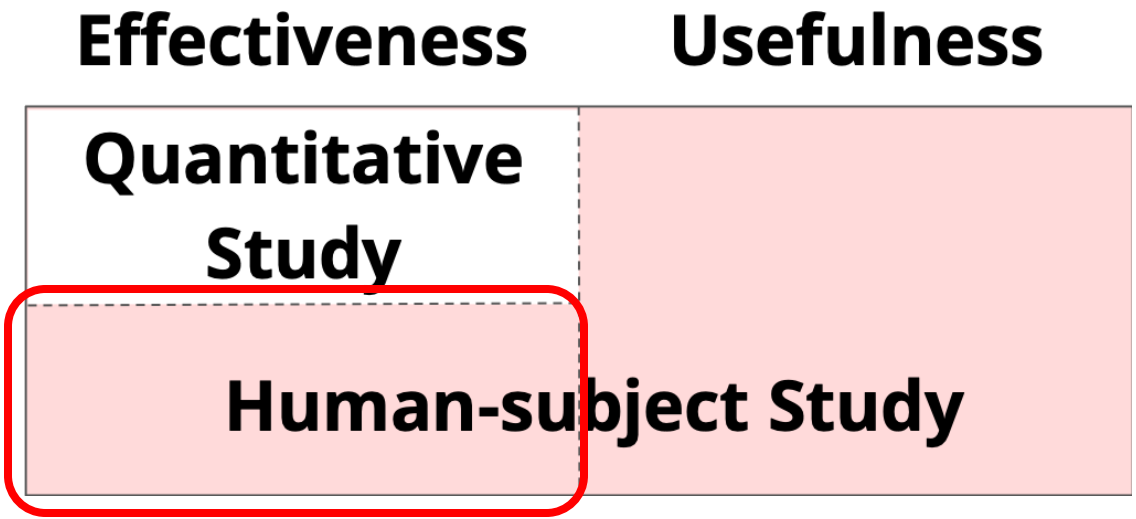
+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods

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



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	ibradypod/phantomjs, last commmit:May 28	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	raff/phantomjs, last commmit: Mar 5	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

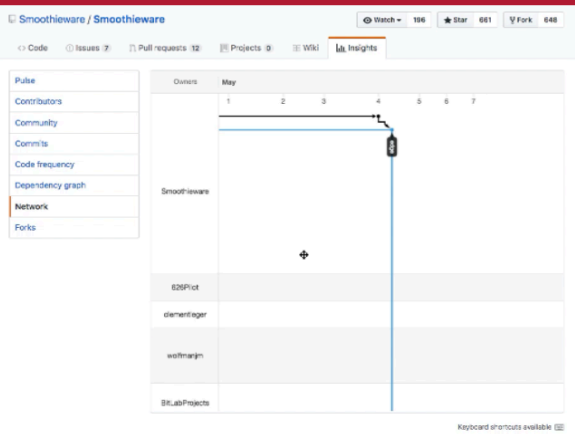
Feature	ricokahler/phantomjs, last commmit:Feb 2	LOC
readlin	readlin, asyncreadrequest, asyncread, qobject, qstring, readline, data, qvariant, m_file, m_data, file, read	30
uint	uint, tmp_value_, value, tmp, octet, qvariant, data, namesize_, readrawdata, fromvalue	80
frame	frame, bmconsumeok, bmdeliver,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

Feature	DeviaVir/phantomjs, last commmit: Jan 25, 2016	LOC
allow	allow, set, customwebpag, ratio, m_customwebpag, devicepixelratio, webpage, setdevicepixelratio	7

VS

Problem

Network View



Human-subject Study - Usefulness

Interesting and Reusable Contribution

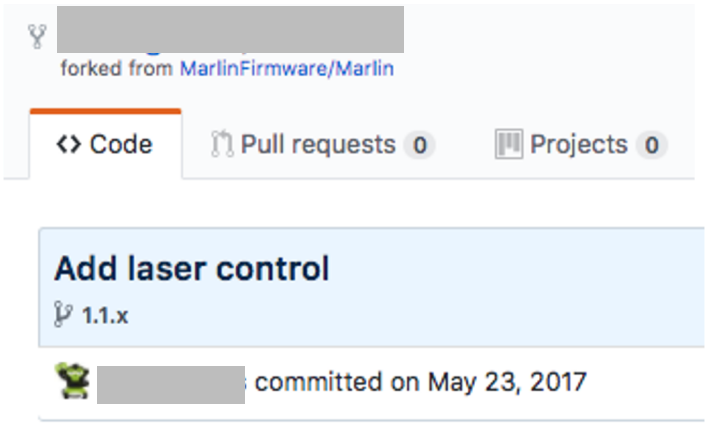
P5: "If it is only exists in this fork, then I want to somehow get this fork into my fork."

Human-subject Study - Usefulness

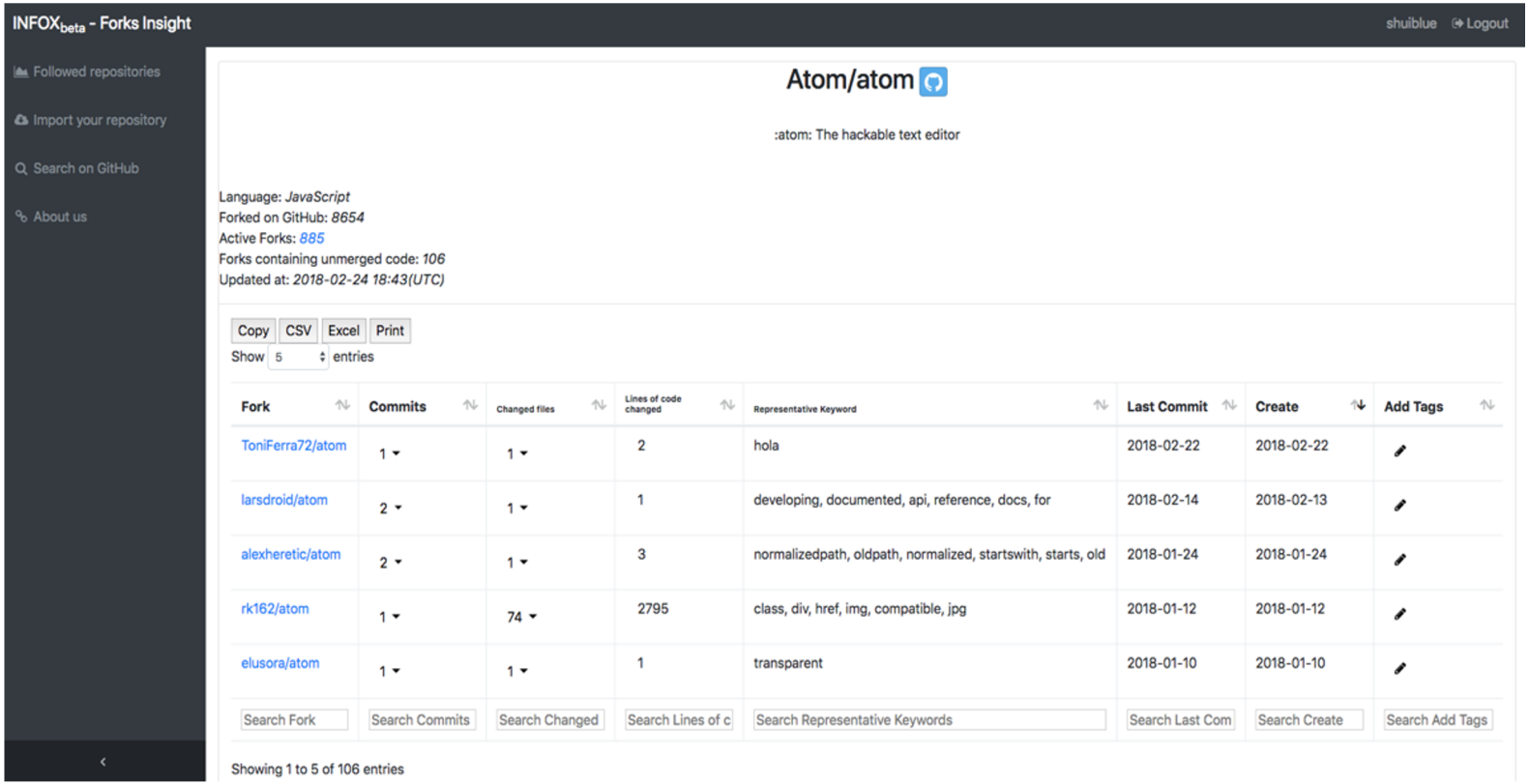
Redundant Development

"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."

P3



forks-insight.com



Problem

Solution

Lack of Overview

Redundant Development

Natural Intervention

Identifying Best Practices

[FSE'19]

[ICSE'20]

New Intervention

[ICSE'18]

Identifying Redundancies

[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 Wąsowski
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:

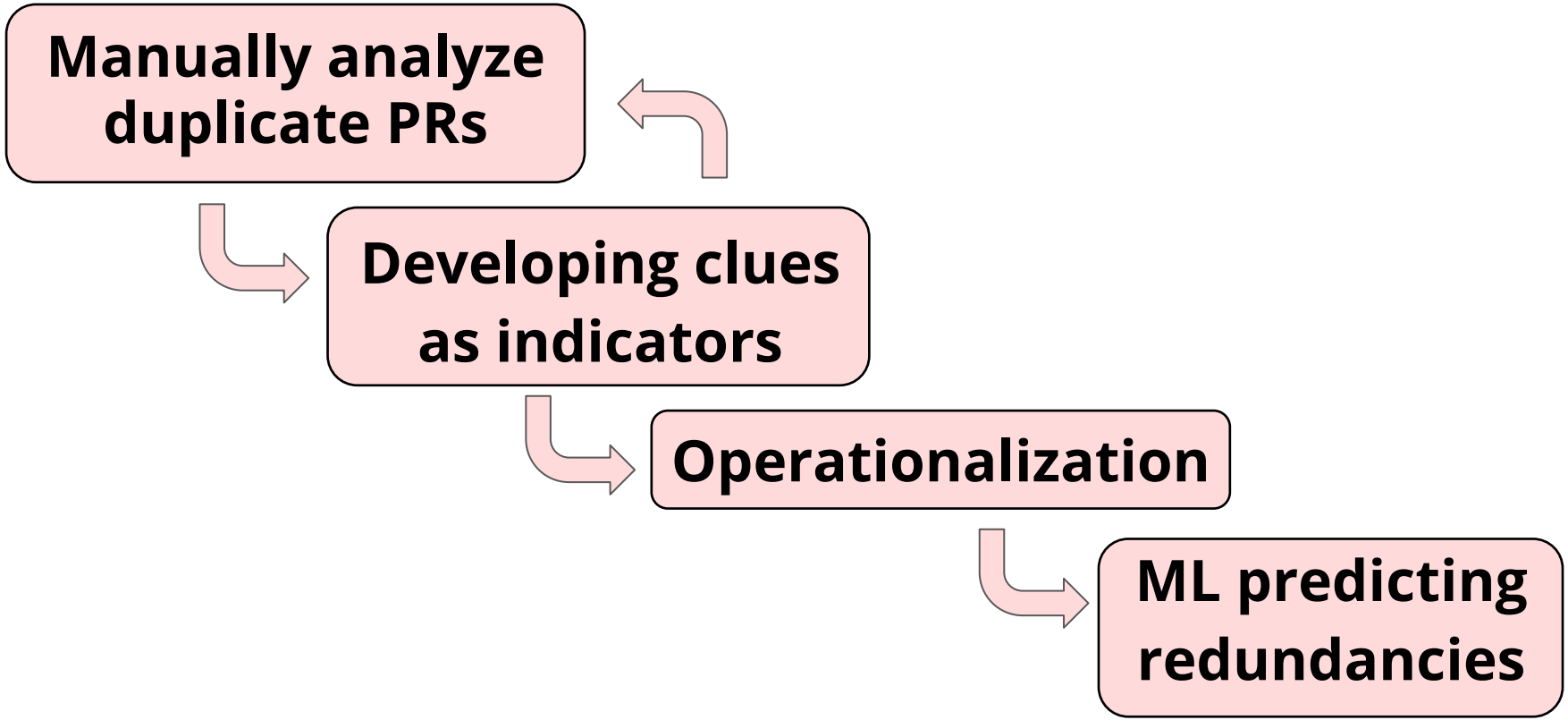
2.6 reviewers

5.2 review comments [Li et al. 2018]

For developers:

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

Research Method

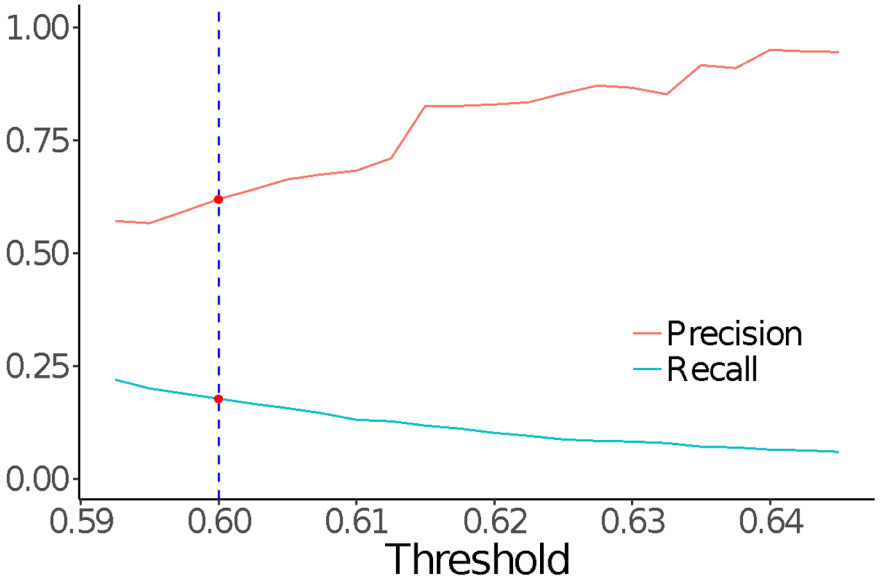


Evaluation - Effectiveness

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?

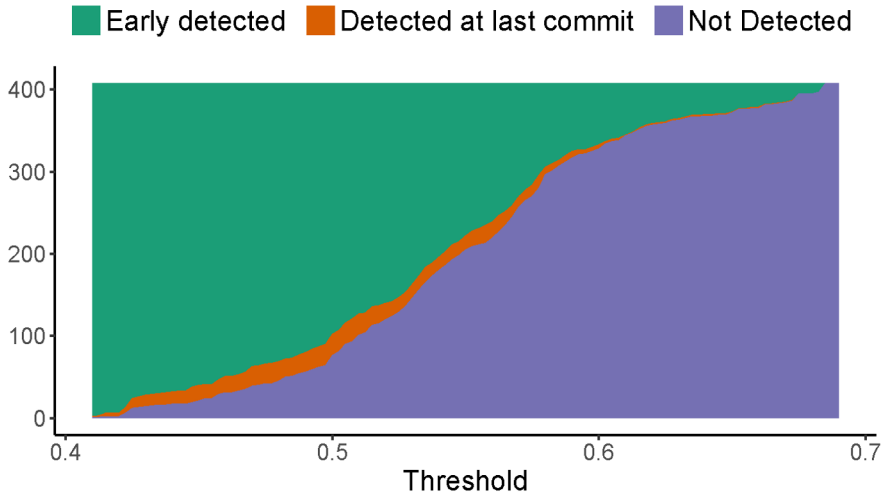
RQ1: helping maintainers to find duplicate PRs



Randomly sample 400 PRs from each project

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

RQ2: helping developers to find duplicate changes early



Recall 46% - 71%

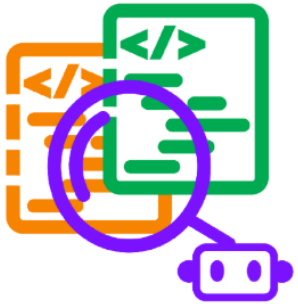
0.07–0.5% false positive rate

Save 1.9 - 3.0 commits per PR

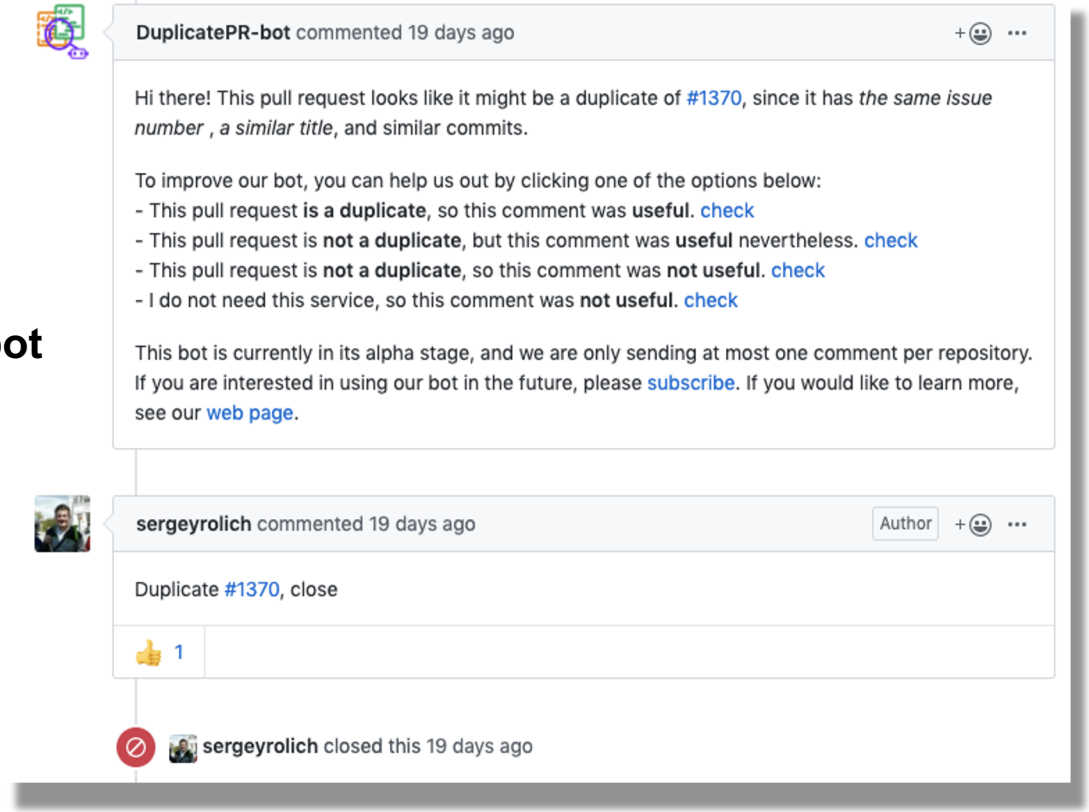
Application Scenario

DuplicatePR-bot

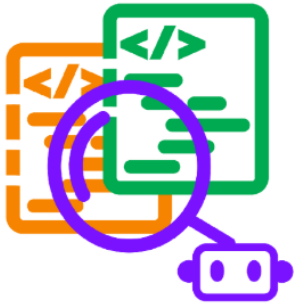
Application Scenario



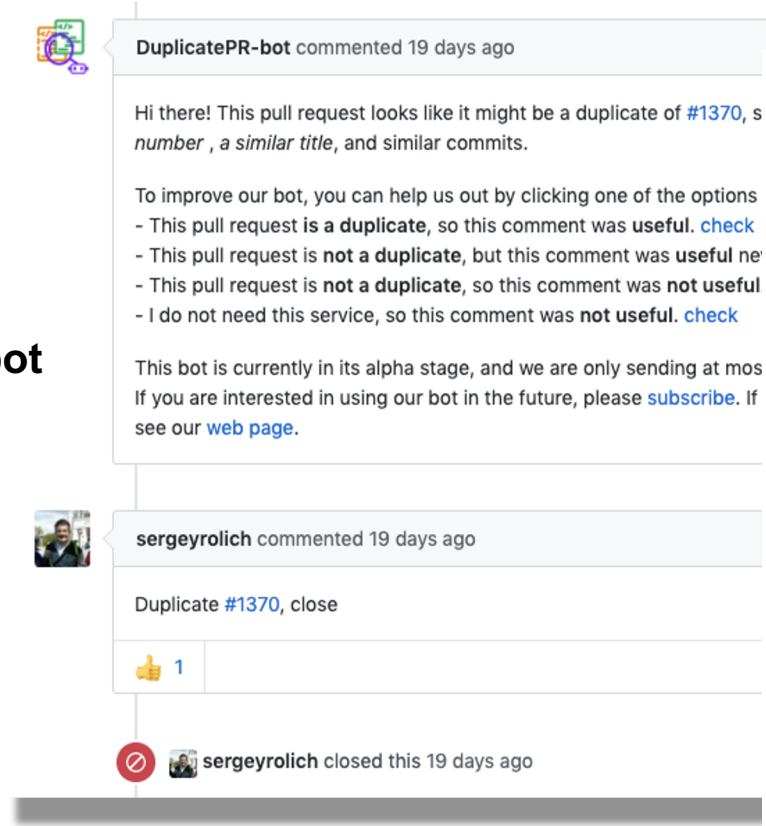
DuplicatePR-bot



Application Scenario



DuplicatePR-bot



Duplicate PR pairs succesfully 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

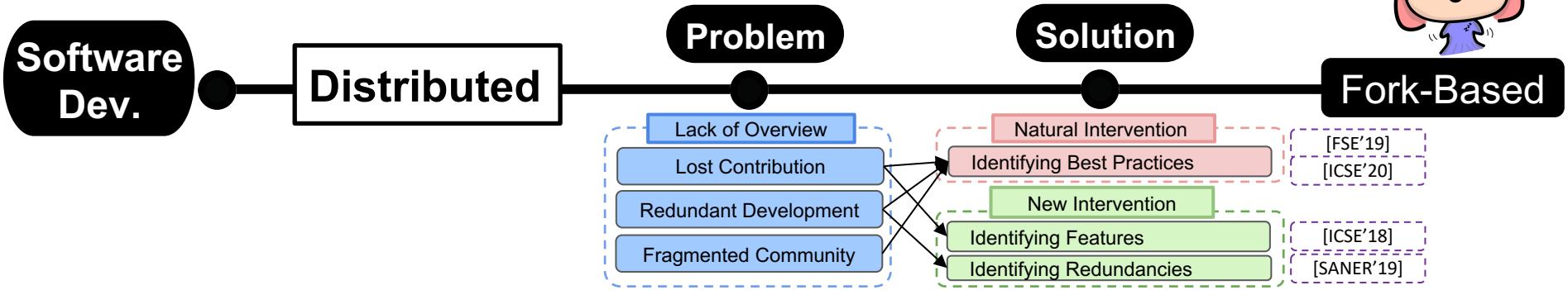
Generalizability

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

Software Dev. — **Distributed** — **Problem** — **Solution** — **Fork-Based**

- Problem
  - Lack of Overview
  - Lost Contribution
  - Redundant Development
  - Fragmented Community
- Solution
  - Natural Intervention
    - Identifying Best Practices — [FSE'19] [ICSE'20]
  - New Intervention
    - Identifying Features — [ICSE'18]
    - Identifying Redundancies — [SANER'19]

FUTURE

Centralized Management Index 98%

Modularity High

Overview of Hard Forks

Improving Collaboration Efficiency

Software Dev. — **Distributed** — **Problem** — **Solution** — **Fork-Based**

Problem

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

Solution

- Natural Intervention
  - Identifying Best Practices — [FSE'19] [ICSE'20]
- New Intervention
  - Identifying Features — [ICSE'18]
  - Identifying Redundancies — [SANER'19]
  - Identifying Lost Contribution
  - Identifying Relevant People
  - Facilitating Particular Interactions

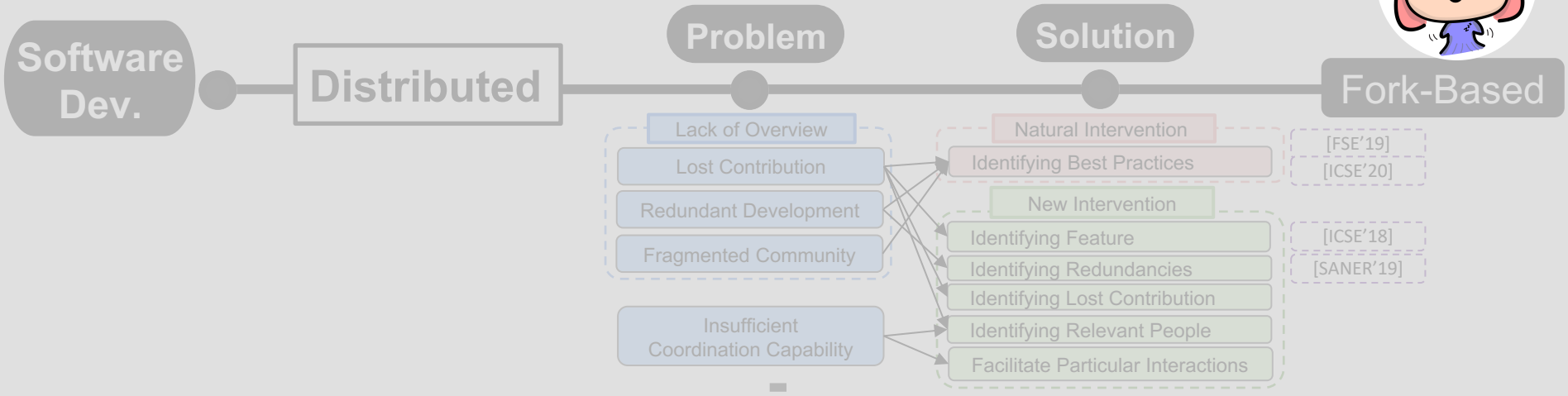
FUTURE

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

Lack of Overview

Lost Contribution

Redundant Development

Fragmented Community

Insufficient Coordination Capability

Natural Intervention

Identifying Best Practices

[FSE'19]

[ICSE'20]

New Intervention

Identifying Feature

Identifying Redundancies

Identifying Lost Contribution

Identifying Relevant People

Facilitate Particular Interactions

[ICSE'18]

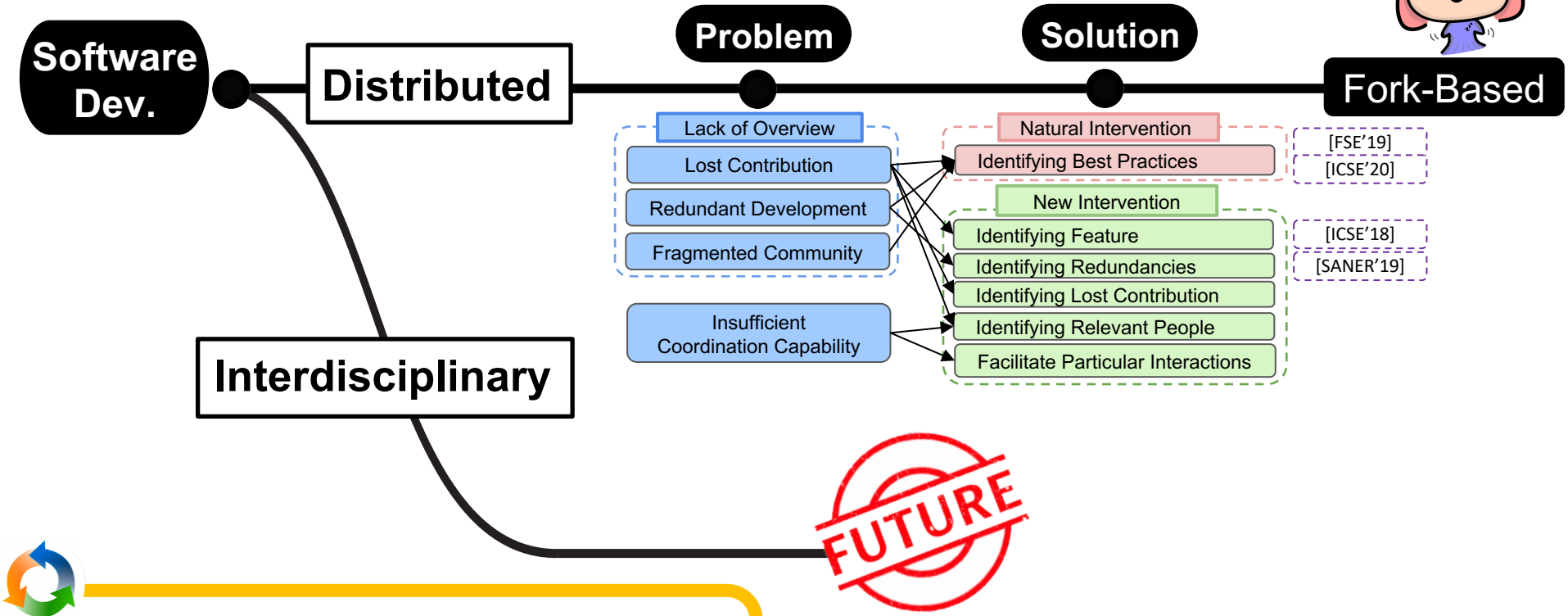
[SANER'19]

Interdisciplinary

FUTURE

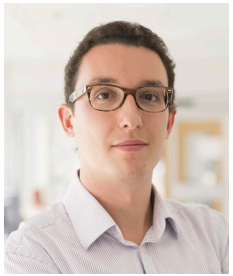
+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods

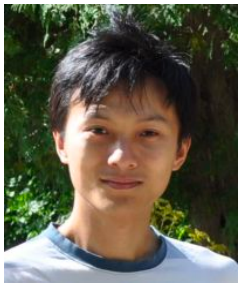
Improving Collaboration Efficiency

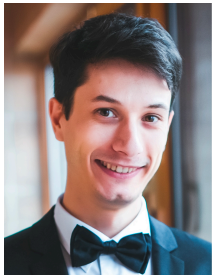


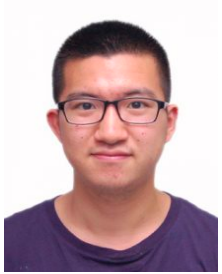
+ Advances in tooling & SE principles
+ Insights from other disciplines
+ Mix a wide range of research methods

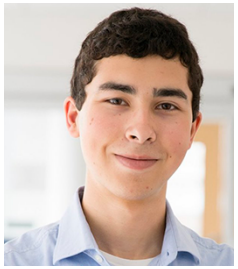
Acknowledgement

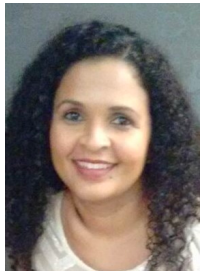












It's not yet the end ...