

論文紹介

We Are Family: Analyzing Communication in GitHub Software Repositories and Their Forks

ソフトウェアファミリ：GitHubリポジトリと フォークのコミュニケーション分析

著者

Scott Brisson, Ehsan Noei, Kelly Lyons

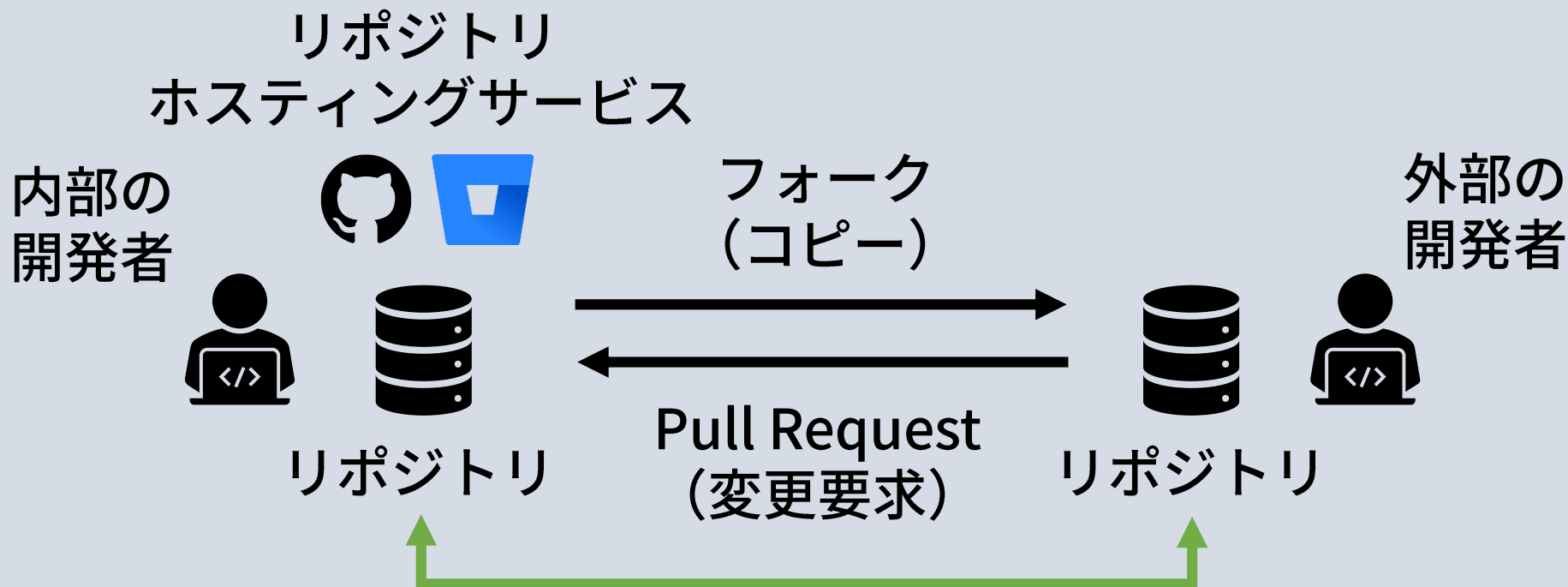
University of Toronto

論文情報：27th IEEE International Conference on Software Analysis, Evolution and Reengineering London, Ontario, Canada, February 18-21, 2020.

紹介者：堀口 日向

1. 背景 & 目的

背景



この研究では、このようなリポジトリ同士の関係を親子関係に見立て、**ソフトウェアファミリ**と呼称

1. 背景 & 目的

背景

フォークは，コミット数やIssue数，人気度など，OSS開発の成功に不可欠な要素と相関関係にある[5]-[7]

しかし，OSSに関する研究の多くは，IssueやPull Requestの投稿がリポジトリ内からなのか，ソフトウェアファミリ内からなのか，さらにその外からなのかといった，投稿者の所属について考慮していない

目的

ソフトウェアファミリ内外からのコミュニケーションとリポジトリの人気（スター数）との関係进行分析

- [5] H. Borges and M. T. Valente, "What's in a github star? understanding repository starring practices in a social coding platform," Journal of Systems and Software, vol. 146, pp. 112–129, 2018.
- [6] T. F. Bissyande, D. Lo, L. Jiang, L. R ´ eveillere, J. Klein, and Y. Le Traon, "Got issues? who cares about it? a large scale investigation of issue trackers from github," in 2013 IEEE 24th international symposium on software reliability engineering (ISSRE). IEEE, 2013, pp. 188–197.
- [7] K. Blincoe, J. Sheoran, S. Goggins, E. Petakovic, and D. Damian, "Understanding the popular users: Following, affiliation influence and leadership on github," Information and Software Technology, vol. 70, pp. 30–39, 2016.

2. 分析に使用するデータセット

GHTorrentに保存されているの約1.1億件のリポジトリ

1. ソフトウェアファミリの抽出

リポジトリのフォーク元の情報から，親子関係を持つリポジトリをソフトウェアファミリとして抽出

2. データの除外

開発者数が少ないリポジトリ，ソフトウェア以外のリポジトリを除外

最終的に，約1.3万件のリポジトリ，
計385件のソフトウェアファミリを抽出

3. 分析方法 & 結果

RQ1. ファミリ内のリポジトリのスター数の分布は？

動機

親リポジトリと子リポジトリでスター数に違いがあるのかを知る

アプローチ

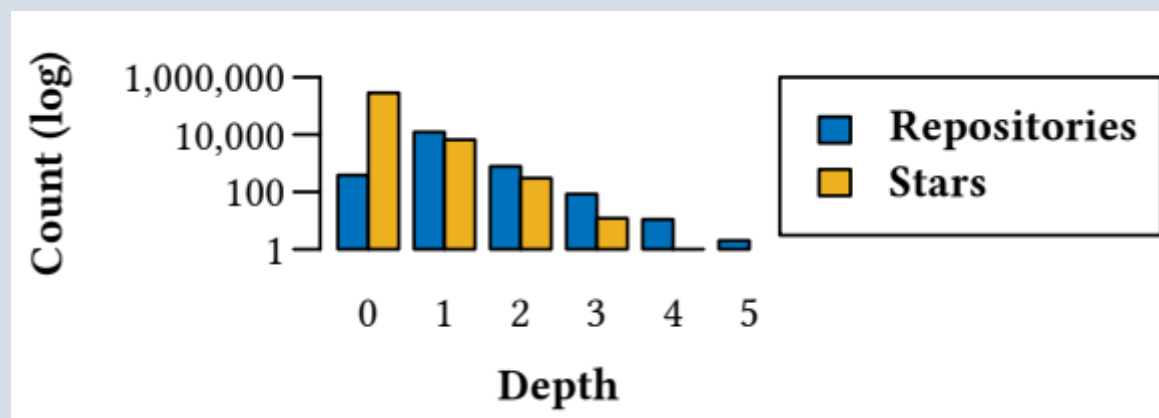
フォークの深さとスター数の分布を，
Dunn法（多重比較検定的一种）を使いファミリ間で比較

3. 分析方法&結果

RQ1. ファミリ内のリポジトリのスター数の分布は？

結果

橙色は
スター数
(対数)



フォークの深さ

- 親リポジトリ（深さ0）のスター数が最も多い（子リポジトリの50倍以上）
- Dunn法により，76%のファミリは同様の分布

3. 分析方法 & 結果

RQ2. リポジトリ内，ファミリー内，ファミリー外からの
コミュニケーション量はどの程度あるのか

動機

投稿元（リポジトリ内，ファミリー内，ファミリー外）によって，
利用するコミュニケーションチャンネルに違いはあるのかを知る

アプローチ

リポジトリ内，ファミリー内，ファミリー外からのIssue投稿数や
Pull Request投稿数などのメトリクスを比較

3. 分析方法 & 結果

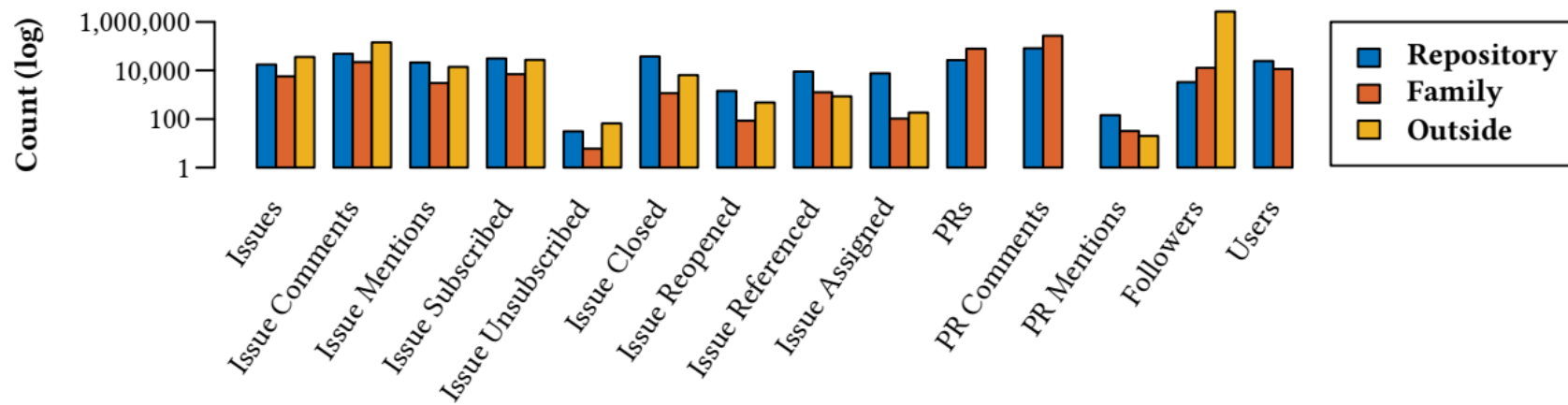
RQ2. リポジトリ内、ファミリー内、ファミリー外からの
コミュニケーション量はどの程度あるのか

結果

- 全体

すべての開発者（24,998人）のうち、約46%（11,499人）が
同じファミリー内の複数のリポジトリに貢献している

件数
(対数)



メトリクス

3. 分析方法&結果

RQ2. リポジトリ内、ファミリー内、ファミリー外からの
コミュニケーション量はどの程度あるのか

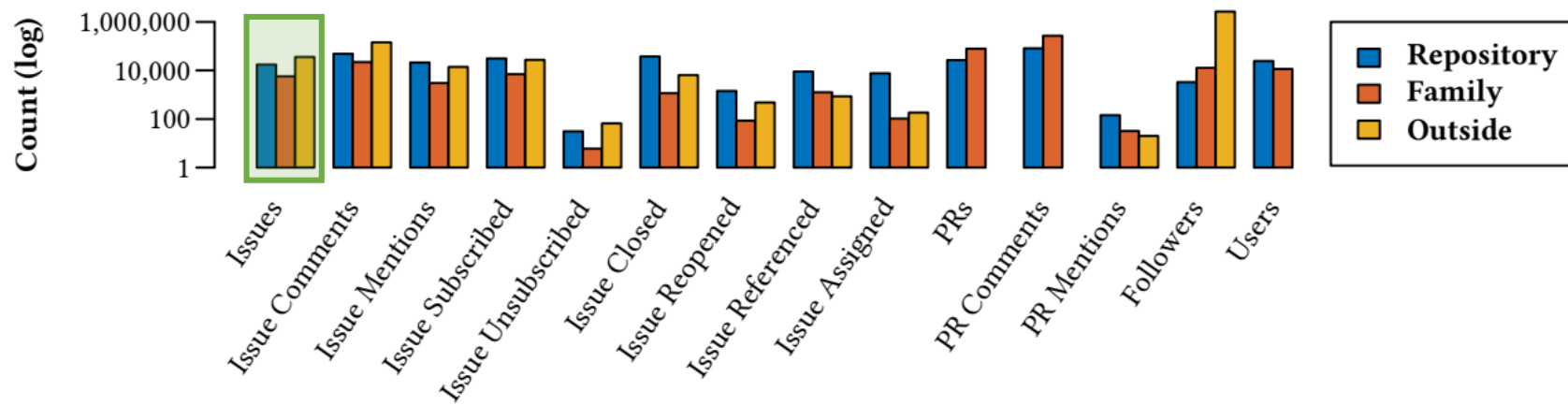
結果 (次以降のRQに関係ある結果だけ説明)

- Issue投稿数の約6万件の内訳

- ・ リポジトリ内からの投稿：約30% (約18,000件)
- ・ ファミリー内からの投稿：約10% (約6,000件)
- ・ ファミリー外からの投稿：約60% (約36,000件)

リポジトリ外からの
Issueが約70%

件数
(対数)



メトリクス

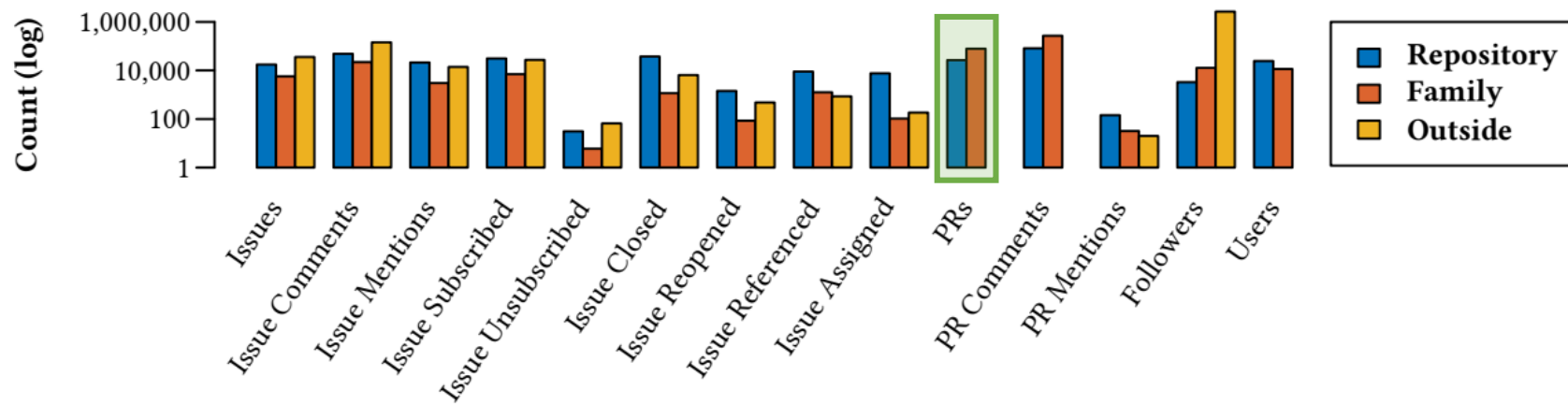
3. 分析方法 & 結果

RQ2. リポジトリ内, ファミリ内, ファミリ外からの コミュニケーション量はどの程度あるのか

結果 (次以降のRQに関係ある結果だけ説明)

- Pull Request投稿数の約6.5万件の内訳
 - リポジトリ内からの投稿: 約40% (26,000件)
 - ファミリ内からの投稿: 約60% (39,000件)

件数
(対数)



メトリクス

3. 分析方法&結果

RQ3. リポジトリ内，ファミリー内，ファミリー外のコミュニケーション量はスター数とどのような関係にあるか？

動機

スター数と関係があるコミュニケーションを明らかにする

アプローチ

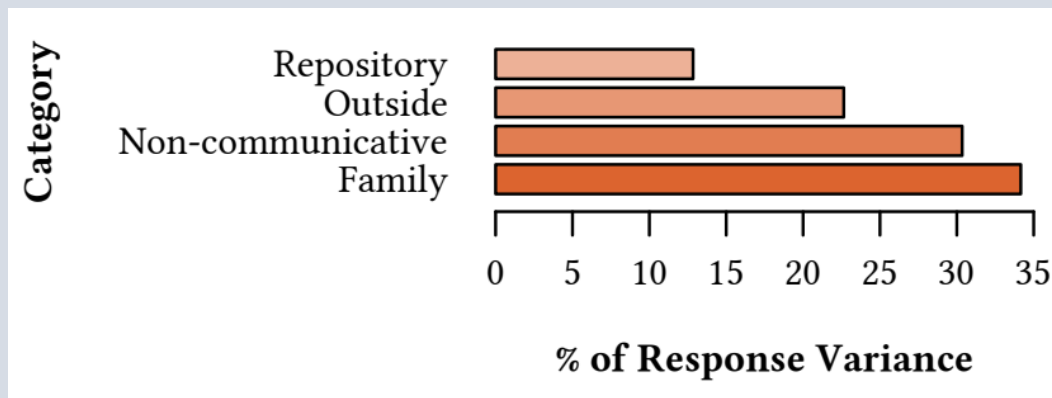
重回帰分析で，リポジトリ内，ファミリー内，ファミリー外のコミュニケーション量とスター数に 관계があるかを調べ，さらに，分散分析で寄与率を調べる

3. 分析方法&結果

RQ3. リポジトリ内，ファミリー内，ファミリー外のコミュニケーション量はスター数とどのような関係にあるか？

結果

投稿元の
カテゴリ



Non-communicative :
リポジトリ自体に
関連するメトリクス
(年数,
フォークの深さ等)

寄与率

- ・ スター数への寄与率が最も高いのはファミリー内からのコミュニケーション
- ・ 外からの影響が1番でないのは意外

3. 分析方法 & 結果

RQ4. リポジトリ内，ファミリー内，ファミリー外でコミュニケーションのトピックに違いはあるか

動機

RQ3の結果を受けて，リポジトリ内，ファミリー内，ファミリー外でスター数に対する寄与率の違う原因を探るため，コミュニケーションの内容を明らかにする

アプローチ

構造的トピックモデル（Structural Topic Model）を使って，Issue約6万件， Pull Request約6.5万件からトピックを抽出

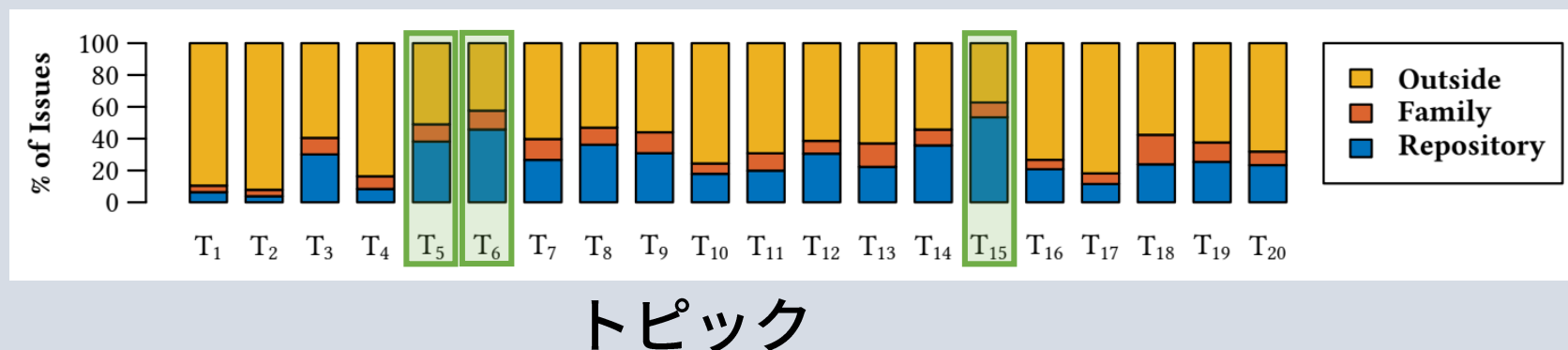
3. 分析方法 & 結果

RQ4. リポジトリ内, ファミリ内, ファミリ外でコミュニケーションのトピックに違いはあるか

結果

• Issueのトピック分析

投稿元の割合



- リポジトリ内からの投稿が多いトピック
Methods (T5) , Features (T6) , Notifications (T15)
➡ ソフトウェアの機能や, リポジトリの管理に関する内容

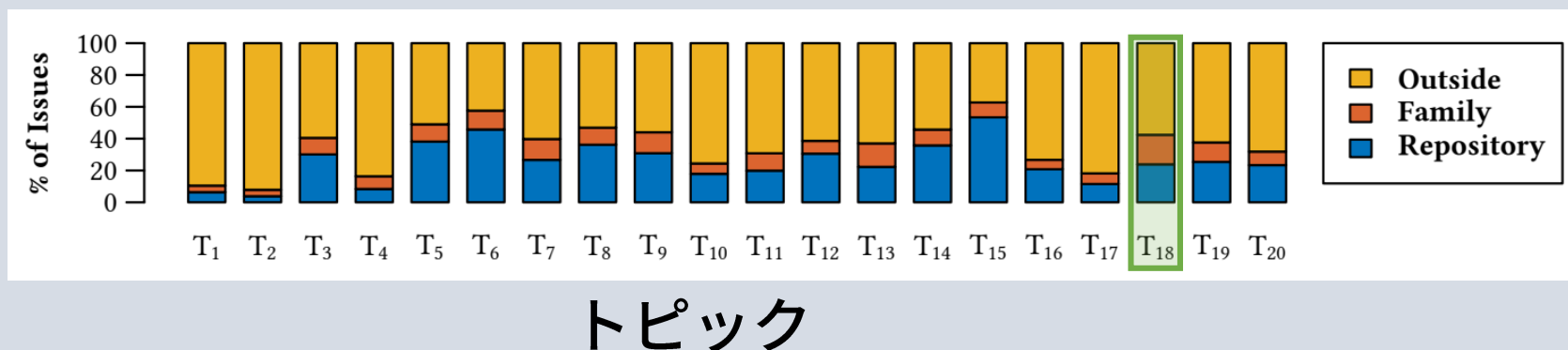
3. 分析方法&結果

RQ4. リポジトリ内, ファミリ内, ファミリ外でコミュニケーションのトピックに違いはあるか

結果

• Issueのトピック分析

投稿元の割合



- ファミリ内からの投稿が多いトピック
Build (T18)

➡ 実際にPull Requestを投稿するために必要な内容

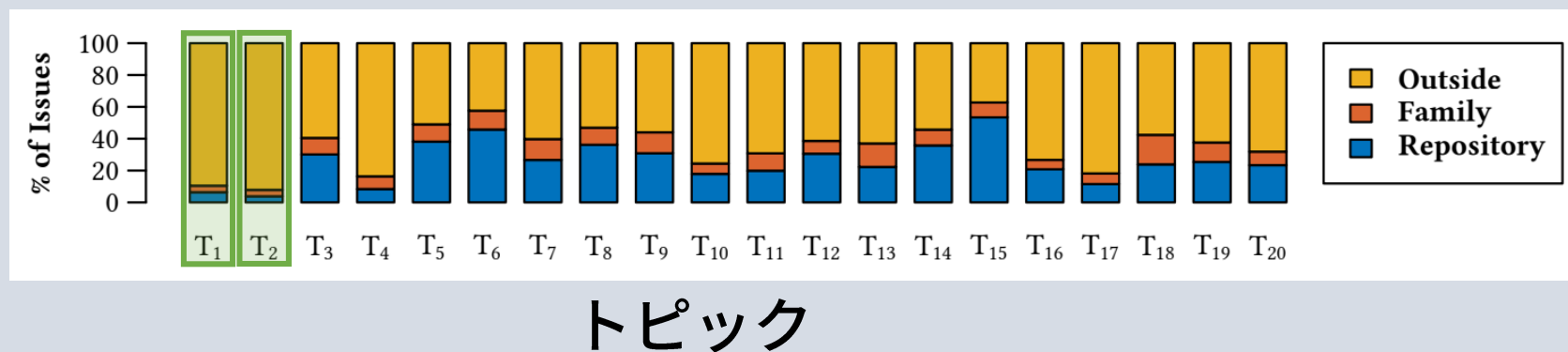
3. 分析方法 & 結果

RQ4. リポジトリ内, ファミリ内, ファミリ外でコミュニケーションのトピックに違いはあるか

結果

• Issueのトピック分析

投稿元の割合



- ファミリ外からの投稿が多いトピック
NextCloud Config (T1) , React UI (T2)

➡ アプリ固有の内容

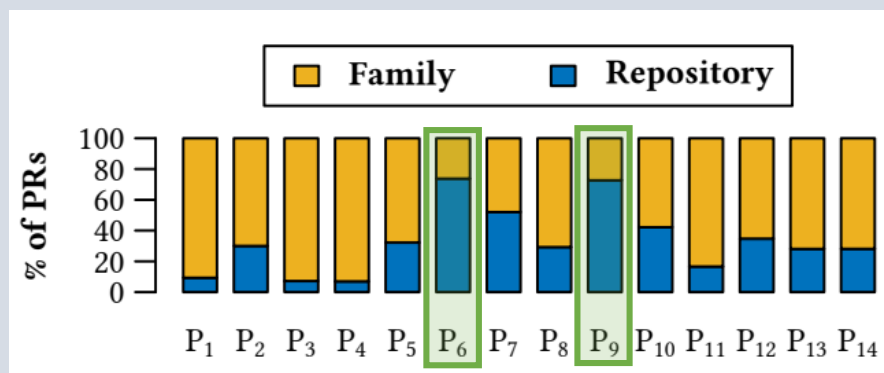
3. 分析方法&結果

RQ4. リポジトリ内, ファミリ内, ファミリ外でコミュニケーションのトピックに違いはあるか

結果

- Pull Requestのトピック分析

投稿元
の割合



トピック

- リポジトリ内からの投稿が多いトピック
Codecov(カバレッジ通知) (P6), Mentionbot/Sync-bot(メンション通知) (P9)



コミュニティやPull Requestの運用・管理に関する内容

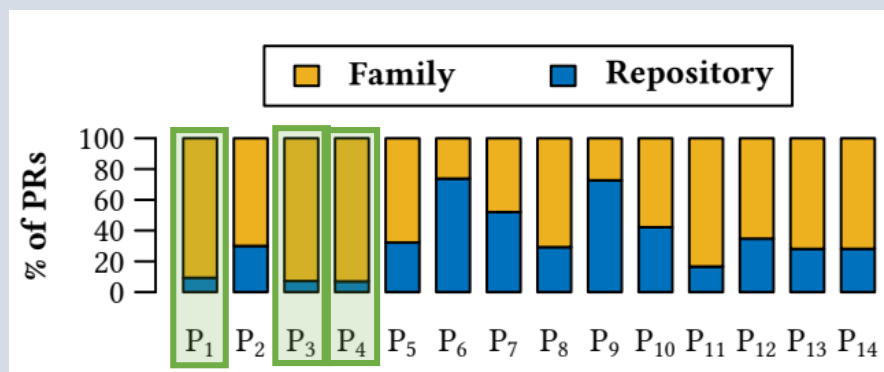
3. 分析方法 & 結果

RQ4. リポジトリ内, ファミリ内, ファミリ外でコミュニケーションのトピックに違いはあるか

結果

- Pull Requestのトピック分析

投稿元
の割合



トピック

- ファミリ内からの投稿が多いトピック

BloomDesktop (P1) , Reviewable (P3) , YCMD (コード補完ツール) (P4)



アプリ固有のトピック, レビューに関するトピック

4. まとめ&所感

結果まとめ

- ・同じソフトウェアファミリー内のユーザ間のコミュニケーションはリポジトリの人気を高めるために重要である
- ・投稿元がリポジトリ内，ファミリー内，ファミリー外のどれかによって，IssueやPull Requestのトピックは異なる

所感

- ・リポジトリの人気には，そのリポジトリだけではなく，フォークしたリポジトリとのコミュニケーションも関係していることを示したのがこの研究のポイント
- ・具体的にソフトウェアファミリーの「何が」リポジトリの人気に影響を及ぼすのかはもっといろいろ調べられそう

補足資料. コミュニケーション量のメトリクス

Category	Metric	Description
Response	stars	Number of stars.
Age	age	How long (in hours) the repository has existed.
Forks	depth	Number of forks away from the parent repository.
	forks	Number of forks created from the repository.
	active_forks	Number of forks created from the repository that opened a PR or issue within the family.
Users	users_repo	Number of users that have write access to the repository.
	users_also_in_family	Number of users that have write access to the repository, and to at least one other repository in the same family.
Followers	followers_repo	Total followers of each user in the same repository.
	followers_family	Total followers of each user in the same family.
	followers_outside	Total followers of each user outside the family.
Pull Requests	pr_repo	Number of PRs that have occurred within the repository (via branching).
	pr_family	Number of PRs that have occurred within the family (via forking).
PR Comments	pr_repo_comments	Number of discussion comments from a PR within the repository.
	pr_family_comments	Number of discussion comments from a PR within the family.
	pr_repo_code_comments	Number of code comments from a PR within the repository.
	pr_family_code_comments	Number of code comments from a PR within the family.
PR Sentiments †	positive_sentiment	The average positive sentiment calculated for each metric in the PR Comments category.
	negative_sentiment	The average negative sentiment calculated for each metric in the PR Comments category.
PR Mentions	pr_mentioned_repo	Number of users mentioned in a PR belonging to the same repository.
	pr_mentioned_family	Number of users mentioned in a PR belonging to the same family.
	pr_mentioned_outside	Number of users mentioned in a PR from outside the family.
	pr_mentioned_unique_repo	Number of unique users mentioned in a PR belonging to the same repository.
	pr_mentioned_unique_family	Number of unique users mentioned in a PR belonging to the same family.
	pr_mentioned_unique_outside	Number of unique users mentioned in a PR from outside the family.
Issues	issue_repo	Number of issues opened by a user belonging to the same repository.
	issue_family	Number of issues opened by a user belonging to the same family.
	issue_outside	Number of issues opened by a user from outside the family.
Issue Comments	issue_comments_repo	Number of comments from issues opened by a user belonging to the same repository.
	issue_comments_family	Number of comments from issues opened by a user belonging to the same family.
	issue_comments_outside	Number of comments from issues opened by a user from outside the family.
Issue Sentiments *	positive_sentiment	The average positive sentiment calculated for each metric in the Issue Comments category.
	negative_sentiment	The average negative sentiment calculated for each metric in the Issue Comments category.
Issue Mentions	issue_mentioned_repo	Number of mentions in an issue opened by a user from the repository.
	issue_mentioned_family	Number of mentions in an issue opened by a user from the family.
	issue_mentioned_outside	Number of mentions in an issue opened by a user outside the family.
	issue_mentioned_unique_repo	Number of unique mentions in an issue opened by a user from the repository.
	issue_mentioned_unique_family	Number of unique mentions in an issue opened by a user from the family.
	issue_mentioned_unique_outside	Number of unique mentions in an issue opened by a user outside the family.
Issue Events *	issue_closed	Number of issues closed.
	issue_subscribed	Number of users who have subscribed to an issue.
	issue_unsubscribed	Number of users who have unsubscribed from an issue.
	issue_reopened	Number of users who have reopened an issue.
	issue_assigned	Number of users assigned to an issue.
	issue_referenced	Number of users referencing an issue.

* All metrics have been calculated for users in the same repository, family, and outside the family.

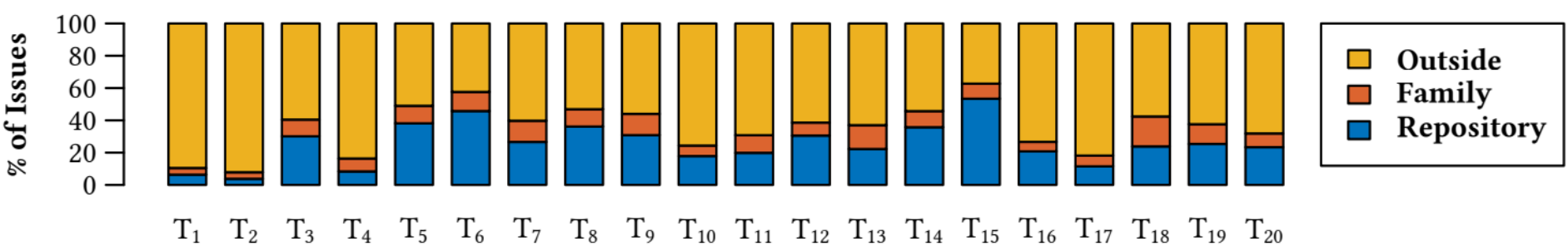
† All metrics have been calculated for users in the same repository and family.

補足資料. 重回帰分析の結果

Metric	Estimate	Pr(>F)	Rel.
Non-communicative	—	—	—
forks	14050.658	< 2e-16 ***	↗
depth	277.189	2.40e-14 ***	↘
age	12.583	0.375914	
Repository	—	—	—
issue_repo	4725.571	< 2e-16 ***	↗
issues_unsubscribed_repo	5693.965	< 2e-16 ***	↘
pr_mentioned_repo	5223.013	5.95e-15 ***	↘
issues_assigned_repo	3460.600	3.41e-14 ***	↘
followers_repo	1539.621	2.24e-12 ***	↘
pr_repo_code_comments	1685.611	1.22e-08 ***	↗
pr_repo_neg	231.079	0.008786 **	↘
pr_repo_comments	4086.671	0.008076 **	↘
pr_repo	2386.168	0.026081 *	↘
pr_repo_pos	40.304	0.613545	
issues_reopened_repo	1399.081	0.238119	
Family	—	—	—
users_also_in_family	1971.051	< 2e-16 ***	↗
pr_mentioned_family	5184.101	< 2e-16 ***	↗
issues_unsubscribed_family	6435.359	< 2e-16 ***	↗
issues_closed_family	7114.712	< 2e-16 ***	↘
issues_reopened_family	5383.633	< 2e-16 ***	↗
issues_assigned_family	2335.077	< 2e-16 ***	↗
issues_referenced_family	2089.421	3.11e-07 ***	↘
pr_family_code_comments	1219.366	0.000627 ***	↘
pr_family	849.382	0.003637 **	↗
pr_family_comments	747.066	0.008731 **	↘
followers_family	130.352	0.018029 *	↘
pr_family_pos	27.776	0.228881	
pr_family_code_neg	11.466	0.776198	
pr_family_neg	3.041	0.899862	
Outside	—	—	—
followers_outside	2441.401	< 2e-16 ***	↗
issues_reopened_outside	9066.449	< 2e-16 ***	↗
issue_outside	6506.352	5.13e-09 ***	↗
issues_unsubscribed_outside	1703.198	7.90e-08 ***	↗
pr_mentioned_outside	1056.548	4.88e-06 ***	↘
issues_referenced_outside	1380.460	0.013112 *	↘
issues_assigned_outside	170.946	0.751762	

補足資料. Issueのトピックと投稿元の内訳

Topic	Topic Name	Top Word Probabilities	Brief Description
T ₁	NextCloud Config.	nextcloud, details, summary, server, removed	Server configuration issues, mainly within the Nextcloud application.
T ₂	React UI	react, screen, navigation, issue, native	User interface issues, mainly with React navigation.
T ₃	Styling	page, file, issue, version, app	General user interface styling.
T ₄	VR and HAS	domoticz, revive, error, occulus, game	Errors mainly with LibreVR and Domoticz.
T ₅	Methods	function, type, string, return, float	Method level compilation, type, and argument errors.
T ₆	Features	time, issue, code, add, support	Suggestions or feature requests.
T ₇	PHP and DB	php, wallabag, table, database, symfony	Hypertext preprocessor and database issues, mainly within the Wallabag application.
T ₈	Presto	error, presto, java, jar, null	Java issues, mainly within the Presto application.
T ₉	Logs	module, cljs, shadow, icinga, modules	Issues including log messages.
T ₁₀	Nextcloud Sharing	www, lib, var, nextcloud, oc	Errors from Nextcloud's sharing functionality.
T ₁₁	Node	npm, node, node_modules, error, js	Issues from Node.js and its package manager npm.
T ₁₂	Build and Installation	version, file, install, docker, run	Issues with building and installing applications, generally from the command line.
T ₁₃	Html and Latex	class, latexml, html, div, text	Display issues with html and latex elements.
T ₁₄	Connectivity	info, debug, error, src, connection	Issues connecting to a service.
T ₁₅	Notifications/BigchainDB	email, notifications, reply, bigchaindb, view	Issues using GitHub's notification service, and BigdbChainDB issues.
T ₁₆	Ruby	lib, gems, logstash, ruby, usr	Issues with Ruby and ruby gems.
T ₁₇	Plugins	plugin, node, jstree, function, android	Issues with plugins, mainly with JSTree plugin.
T ₁₈	Build	build, src, include, usr, error	Issues with building applications.
T ₁₉	Python	python, line, file, lib, packages	Issues with applications using Python.
T ₂₀	HTTP	server, user, client, error, http	Connectivity, requests, and routing issues.



補足資料. PRのトピックと投稿元の内訳

Topic	Topic Name	Top Word Probabilities	Brief Description
P_1	BloomDesktop	src, bloomdesktop, bloombooks, file, reviews	PRs for the BloomDesktop application.
P_2	Additions	add, type, method, function, code	Adding new functionality.
P_3	Reviewable	files, reviewable, reviewed, img, alt	PRs using Reviewable, a code review application.
P_4	YCMD	ycmd, valloic, reviews, file, line, completers	PRs surrounding YCMD, a code-completion server.
P_5	Google Bot and PRs	pull, branch, request, git, version	Google bot, or PRs talking about GitHub's functionality such as PRs, forking, merging, and branching.
P_6	Codecov	src, diff, pr, el, tree	PRs using Codecov.io, a code-coverage application.
P_7	React CI and Benchmarking	php, channel, mb, react, ms	React-navigation-ci and speed benchmarking.
P_8	Coveralls	coverage, status, badge, pulling, code	PRs using Coveralls, a testing code coverage application.
P_9	Mentionbot/Snyk-bot	user, files, pr, request, app	PRs using Mentionbot and Snyk-bot.
P_{10}	Server Time & Logs	code, time, patch, test, server	PRs surrounding time (delays, timeouts, rates etc.) and logging, mainly with servers.
P_{11}	Pentaho	pentaho, java, src, org, osgi	PRs with Pentaho's Open Source OSGI bundles.
P_{12}	Tables	fix, test, update, tests, add	PRs that use a checklist, table, or common format.
P_{13}	Builds	file, build, pr, docker, windows	PRs mentioning builds.
P_{14}	Python	python, pr, error, it, gt	PRs discussing updating Python dependencies.

