

Unveiling Elite Developers' Activities in Open Source Project

ZHENDONG WANG* , University of California, Irvine

YANG FENG* , University of California, Irvine

YI WANG, Rochester Institute of Technology

JAMES A. JONES, University of California, Irvine

DAVID REDMILES, University of California, Irvine

ESEC/FSE 2020

紹介者 大森伊月

背景・目的

- 背景

- OSSは重要なインフラである
- OSS開発において大半のコードを書き，進捗管理を行っているのは一部のエリート開発者である
- エリート開発者の活動を分析することはOSSプロジェクトの発展を理解する上で重要

- 目的

- エリート開発者の活動がプロジェクトの生産性や品質などにどのような影響を与えるかを明らかにする

研究課題

- RQ1：OSSプロジェクトにおいてエリート開発者は典型的なコーディング以外に何をしているのか？
- RQ2：OSSプロジェクトにおいてエリート開発者の労力の配分は、プロジェクトの発展に伴ってどのように変化するのか？
- RQ3：OSSプロジェクトにおいてエリート開発者の労力の配分とプロジェクトの成果(生産性と品質)の間にどのような関係があるのか？

アプローチ1

- ・ プロジェクトの成果である生産性と品質を定量的に分析する
 - ・ 生産性
 - ・ あるプロジェクト i のある月 m に行われた全ての新規コミットの数 $NewC_{im}$
 - ・ あるプロジェクト i のクローズドバグのある月 m の平均タイムサイクル BCT_{im}
 - ・ 品質
 - ・ あるプロジェクト i のある月 m に発見された新しいバグ $NewB_{im}$
 - ・ あるプロジェクト i のある月 m におけるバグ修正率 (バグ修正数/バグ発見数) BFR_{im}

アプローチ 2

- GitHubにおけるアクションを4つのアクティビティカテゴリに分類
 - コミュニケーション
 - 課題, コミット, マイルストーンについての公開されたコミュニケーション
 - 組織的活動
 - コミュニティを管理し, コードレビュー・デバッグなどのタスクをコントリビュータにアサインすること
 - サポート
 - ドキュメント作成・バージョン管理などの非コーディング作業
 - 典型的活動
 - 個人レベルで可能なコード作成タスク

アプローチ 3

- RQ1
 - 記述統計
- RQ2
 - 記述統計, 分散分析法
- RQ3
 - プロジェクト別固定効果パネル回帰

データセット

- GitHub上から20のリポジトリを以下の点から選定し，分析する
 - 大規模プロジェクトであり，管理体制が確立されていること
 - アプリケーションドメインの観点から多様なサンプルであること
 - 企業がスポンサーとなっているプロジェクトを含められていること
 - GitHubで追跡可能であること
- エリート開発者の特定
 - オーナーから権限を与えられていて，リポジトリに対して書き込み許可を得ている人
 - 権限を与えられているか分からないので，権限が必要なアクションをおこなったユーザーをリストアップし特定した

RQ1：OSSプロジェクトにおいてエリート開発者は典型的なコーディング以外に何をしているのか？

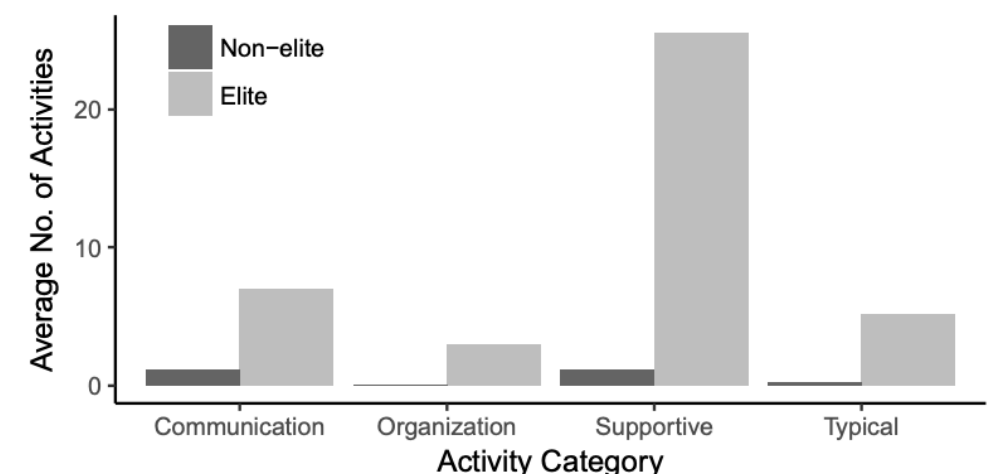
- エリート開発者のアクティビティログがどのアクティビティカテゴリに分類されるかを確認する

- 結果

- 60%以上のエリート開発者がコード貢献を行うだけでなく、ドキュメントを作成したり、議論に参加するなど様々な形でプロジェクトに貢献している
- エリート開発者とそうでない開発者を比較すると全てのカテゴリにおいて月間平均活動数が非常に多くなっている

Table 4. Activities amount that has happened on each sampled project.

Project	Com.		Org.		Sup.		Typ.	
	Total	Elite%	Total	Elite%	Total	Elite%	Total	Elite%
Aframe	4908	0.46	455	0.99	19400	0.85	5180	0.72
Alamofire	5967	0.18	1773	1.00	11906	0.61	1465	0.62
Exoplayer	9293	0.32	2293	1.00	22197	0.71	5361	0.87
finagle	2488	0.30	46	0.93	2947	0.46	2753	0.49
fresco	5283	0.29	290	1.00	10481	0.64	1923	0.77
guava	3161	0.29	664	0.99	7724	0.71	2239	0.53
immutable-js	2909	0.15	28	0.75	5869	0.59	1057	0.52
jest	19073	0.36	1025	0.99	39995	0.66	5015	0.43
marko	1937	0.38	403	0.95	5525	0.79	2956	0.93
Moya	5376	0.43	416	0.61	22808	0.42	2860	0.73
nightmare	3105	0.14	24	1.00	4963	0.48	892	0.50
rclone	7475	0.23	182	1.00	15243	0.66	2781	0.80
react	35086	0.37	3730	1.00	83036	0.74	9640	0.59
recharts	3199	0.15	54	0.85	4980	0.41	1396	0.66
splitebrowser	6129	0.42	493	1.00	11589	0.70	1751	0.84
stf	1694	0.33	25	0.64	2672	0.55	837	0.69
tensorflow	97940	0.48	28236	0.92	183485	0.75	43029	0.50
tesseract	4870	0.35	116	1.00	9805	0.59	2512	0.55
tidb	16240	0.80	1944	0.98	45451	0.91	8305	0.89
ZeroNet	2853	0.27	120	1.00	4881	0.56	2650	0.79
Mean	11949.30	0.34	2115.85	0.93	25747.85	0.64	5230.10	0.67



RQ2：OSSプロジェクトにおいてエリート開発者の労力の配分は、プロジェクトの発展に伴ってどのように変化するのか？

- 最も急な成長を遂げたプロジェクトであるTenssorrowflowで、各アクティビティカテゴリの数を分析した
- 結果
 - プロジェクトの発展に伴ってコミュニケーション・サポート・組織的活動のイベントは着実に増加していると言える
 - エリート開発者によるコード投稿量は安定していることがわかる

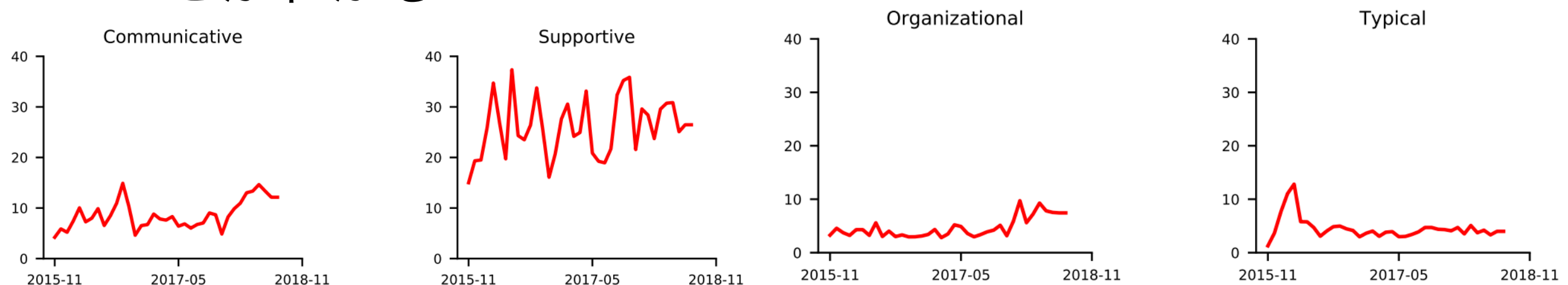


Fig. 5. Trends of individual elite developer's activities in the four activity categories of the Tensorflow.

RQ2：OSSプロジェクトにおいてエリート開発者の労力の配分は、プロジェクトの発展に伴ってどのように変化するのか？

- あるプロジェクト*i*のある月*m*におけるエリート開発者一人当たりの活動の月平均成長率を先ほど増加傾向にあったカテゴリに分けて算出し、一時配置分散分析をおこなった
- 結果
 - ・ 有意な差が見られた($F_{(2,57)} = 8.452, p < 0.001$)
 - ・ それぞれのカテゴリ間の違いを明らかにする必要がある

RQ2：OSSプロジェクトにおいてエリート開発者の労力の配分は、プロジェクトの発展に伴ってどのように変化するのか？

- 3つのカテゴリ間の成長率の違いを明らかにするために Tukey's HSD検定を用いたポストホック分析を行う
- 結果
 - 典型的活動の成長率が他の2つのカテゴリの成長率より有意に低いことがわかった
- RQ2の結果
 - エリート開発者はプロジェクトの発展に伴い、コミュニケーションやサポートに力を入れる傾向にあり、典型的活動への関与は大幅に減少する

RQ3：OSSプロジェクトにおいてエリート開発者の労力の配分とプロジェクトの成果(生産性と品質)の間にどのような関係があるのか？

- エリート開発者一人当たりの4つのカテゴリの活動の中でシェアを占める割合をそれぞれ次のように表す

$$\overline{S - Com_{im}}, \overline{S - Org_{im}}, \overline{S - Sup_{im}}, \overline{S - Type_{im}}$$

- これらの4つの変数を独立変数とし, $NewC_{im}, BCT_{im}, NewB_{im}, BFR_{im}$ を従属変数としてこれらの関係性を固定効果パネル回帰を用いて明らかにする

$$(NewC_{im} = \beta_1 \times \overline{S - Com_{im}} + \beta_2 \times \overline{S - Org_{im}} + \beta_3 \times \overline{S - Sup_{im}} + \alpha_i + u_{it})$$

RQ3：OSSプロジェクトにおいてエリート開発者の労力の配分とプロジェクトの成果(生産性と品質)の間にどのような関係があるのか？

- 生産性の回帰結果
 - モデルP1においてはComとSupで有意であり，ともに負の相関がある
 - モデルP2においてはOrgとSupで有意であり，Orgは負の相関・Supは正の相関がある

Table 5. Regression models for project productivity.

	Whole Sample		Sub-sample (Non-Company)		Sub-sample (Company)	
	New Commit	Bug Cycle Time	New Commit	Bug Cycle Time	New Commit	Bug Cycle Time
	Model P1 (β)	Model P2 (β)	Model P3 (β)	Model P4 (β)	Model P5 (β)	Model P6 (β)
	(SE)	(SE)	(SE)	(SE)	(SE)	(SE)
$S - Com_{im}$	-155.96** (49.20)	-170.71 (143.96)	-125.07*** (28.11)	-381.80 (212.09)	-228.68* (92.85)	-16.44 (202.66)
$S - Org_{im}$	1.38 (121.25)	-148.61* (54.65)	-137.28* (70.24)	-214.71* (129.6)	203.10 (223.47)	-153.85 (187.49)
$S - Sup_{im}$	-138.21*** (35.52)	473.60*** (103.94)	-91.25*** (20.34)	553.17** (153.39)	-204.66** (66.16)	401.30** (144.40)
Unobserved time-invariant effects (α_i) [¶]	-.***	-.***	-.***	-.***	-.***	-.***
Multiple R^2	0.884	0.745	0.686	0.740	0.891	0.751
Adjusted Multiple R^2	0.880	0.736	0.673	0.730	0.887	0.742
$F^\dagger$	231.10***	88.35***	60.45***	73.89	222.00***	82.44***

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

[¶]: The unobserved time-invariant effects are a vector rather than a single coefficient. To keep the paper concise, we do not include them, instead, we show the significant levels of them.

[†]: For Model 1 – 2: degrees of freedom (DFs) are (23, 697); for Model 3 – 4: degrees of freedom are (12, 312); for Model 5 – 6: degrees of freedom are (14, 382).

RQ3：OSSプロジェクトにおいてエリート開発者の労力の配分とプロジェクトの成果(生産性と品質)の間にどのような関係があるのか？

- 生産性の回帰結果
 - モデルP3~P6についてはP1,P2とほぼ同じであり，若干の違いとしてP3ではOrgが有意であり，P6ではOrgが有意でなくなっている
 - P5とP6の R^2 が他とくらべ20%以上高くなっている

Table 5. Regression models for project productivity.

	Whole Sample		Sub-sample (Non-Company)		Sub-sample (Company)	
	New Commit Model P1 (β) (SE)	Bug Cycle Time Model P2 (β) (SE)	New Commit Model P3 (β) (SE)	Bug Cycle Time Model P4 (β) (SE)	New Commit Model P5 (β) (SE)	Bug Cycle Time Model P6 (β) (SE)
$S - Com_{im}$	-155.96** (49.20)	-170.71 (143.96)	-125.07*** (28.11)	-381.80 (212.09)	-228.68* (92.85)	-16.44 (202.66)
$S - Org_{im}$	1.38 (121.25)	-148.61* (54.65)	-137.28* (70.24)	-214.71* (129.6)	203.10 (223.47)	-153.85 (187.49)
$S - Sup_{im}$	-138.21*** (35.52)	473.60*** (103.94)	-91.25*** (20.34)	553.17** (153.39)	-204.66** (66.16)	401.30** (144.40)
Unobserved time-invariant effects (α_i) [¶]	-.***	-.***	-.***	-.***	-.***	-.***
Multiple R^2	0.884	0.745	0.686	0.740	0.891	0.751
Adjusted Multiple R^2	0.880	0.736	0.673	0.730	0.887	0.742
$F^\dagger$	231.10***	88.35***	60.45***	73.89	222.00***	82.44***

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

[¶]: The unobserved time-invariant effects are a vector rather than a single coefficient. To keep the paper concise, we do not include them, instead, we show the significant levels of them.

[†]: For Model 1 – 2: degrees of freedom (DFs) are (23, 697); for Model 3 – 4: degrees of freedom are (12, 312); for Model 5 – 6: degrees of freedom are (14, 382).

RQ3：OSSプロジェクトにおいてエリート開発者の労力の配分とプロジェクトの成果(生産性と品質)の間にどのような関係があるのか？

- 生産性の回帰結果からわかること
 - 組織的活動に積極的に参加することでエリート開発者がバグ修正に時間を割くことになりBCTの長期化につながる可能性がある
 - 同様にサポートに関してもBCTの長期化につながる可能性がある
 - 企業が主催するプロジェクトでは日常的なバグ修正のプロセスが確立されていることが多いのでエリート開発者がそれほど重要でない可能性がある

RQ3：OSSプロジェクトにおいてエリート開発者の労力の配分とプロジェクトの成果(生産性と品質)の間にどのような関係があるのか？

品質の回帰結果

- モデルQ1におけるOrgとSupが有意であり，それぞれ正の相関がある
- モデルQ2におけるComとSupが有意であり，Comは負の相関・Supは正の相関がある

Table 6. Regression models for project quality.

	Whole Sample		Sub-sample (Non-Company)		Sub-sample (Company)	
	New Bug	Bug Fix Rate	New Bug	Bug Fix Rate	New Bug	Bug Fix Rate
	Model Q1 (β)	Model Q2 (β)	Model Q3 (β)	Model Q4 (β)	Model Q5 (β)	Model Q6 (β)
	(SE)	(SE)	(SE)	(SE)	(SE)	(SE)
$S - Com_{im}$	4.13 (5.26)	-2.24*** (0.35)	1.43 (3.91)	-1.28*** (0.34)	5.23 (9.59)	-3.11*** (0.61)
$S - Org_{im}$	50.13*** (12.97)	-0.63 (0.87)	6.48 (9.77)	-0.62 (0.85)	88.69*** (23.09)	-0.35 (1.47)
$S - Sup_{im}$	18.31*** (3.80)	1.12*** (0.26)	-7.67** (2.83)	0.60* (0.25)	26.99*** (6.84)	1.50*** (0.44)
Unobserved time-invariant effects (α_i) [¶]	-.***	-.***	-.***	-.***	-.***	-.***
Multiple R^2	0.857	0.649	0.667	0.761	0.871	0.608
Adjusted Multiple R^2	0.853	0.637	0.654	0.752	0.866	0.594
$F^\dagger$	186.2***	56.09***	52.05***	83.07***	184.2***	42.31***

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

[¶]: The unobserved time-invariant effects are a vector rather than a single coefficient. To keep the paper concise, we do not include them, instead, we show the significant levels of them.

[†]: For Model 1 – 2: degrees of freedom (DFs) are (23, 697); for Model 3 – 4: degrees of freedom are (12, 312); for Model 5 – 6: degrees of freedom are (14, 382).

RQ3：OSSプロジェクトにおいてエリート開発者の労力の配分とプロジェクトの成果(生産性と品質)の間にどのような関係があるのか？

- 品質の回帰結果
 - モデルQ3~Q6においては生産性同様にQ1とQ2とほぼ変わらない結果となった
 - 違いとしてQ3においてOrgの有意がなくなった

Table 6. Regression models for project quality.

	Whole Sample		Sub-sample (Non-Company)		Sub-sample (Company)	
	New Bug	Bug Fix Rate	New Bug	Bug Fix Rate	New Bug	Bug Fix Rate
	Model Q1 (β)	Model Q2 (β)	Model Q3 (β)	Model Q4 (β)	Model Q5 (β)	Model Q6 (β)
	(SE)	(SE)	(SE)	(SE)	(SE)	(SE)
$\overline{S - Com_{im}}$	4.13 (5.26)	-2.24*** (0.35)	1.43 (3.91)	-1.28*** (0.34)	5.23 (9.59)	-3.11*** (0.61)
$\overline{S - Org_{im}}$	50.13*** (12.97)	-0.63 (0.87)	6.48 (9.77)	-0.62 (0.85)	88.69*** (23.09)	-0.35 (1.47)
$\overline{S - Sup_{im}}$	18.31*** (3.80)	1.12*** (0.26)	-7.67** (2.83)	0.60* (0.25)	26.99*** (6.84)	1.50*** (0.44)
Unobserved time-invariant effects (α_i) [¶]	-.***	-.***	-.***	-.***	-.***	-.***
Multiple R^2	0.857	0.649	0.667	0.761	0.871	0.608
Adjusted Multiple R^2	0.853	0.637	0.654	0.752	0.866	0.594
$F^\dagger$	186.2***	56.09***	52.05***	83.07***	184.2***	42.31***

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

[¶]: The unobserved time-invariant effects are a vector rather than a single coefficient. To keep the paper concise, we do not include them, instead, we show the significant levels of them.

[†]: For Model 1 – 2: degrees of freedom (DFs) are (23, 697); for Model 3 – 4: degrees of freedom are (12, 312); for Model 5 – 6: degrees of freedom are (14, 382).

RQ3：OSSプロジェクトにおいてエリート開発者の労力の配分とプロジェクトの成果(生産性と品質)の間にどのような関係があるのか？

- 品質の回帰結果からわかること
 - 非技術的な仕事をするとうリート開発者はコードに取り組む時間が少なくなる可能性があり，それにより非エリート開発者のコード開発に労力を割く必要がありより多くのバグを含んでいる可能性がある
 - サポートにより多くの労力を割くことでバグ除去プロセスを機能させることができる可能性がある

結論

- ・ エリート開発者はプロジェクトの活動の大部分に従事していることはプロジェクトの長期的な目線では最適でない可能性がある
- ・ エリート開発者は次第にサポート業務などに時間を費やすことになり、これが開発者の意図したキャリアではない可能性がある
- ・ プロジェクトの発展に伴ってエリート開発者は非技術的な作業が多くなり、生産性と品質の低下を招くことになるが、これはプロジェクトの作業プロセスを維持するのに役立ち品質の向上につながる
- ・ 企業がスポンサーとなっているプロジェクトではエリート開発者の労力配分による影響を強く受ける

まとめ

- 目的
 - OSSにおいてエリート開発者がどのような役割を果たしているかを調査する
- アプローチ
 - 記述統計, 分散分析法, プロジェクト別固定効果パネル回帰などを用いた定量的な調査
- 結論
 - エリート開発者は様々な活動に参加しているが技術的貢献は一部である
 - エリート開発者はプロジェクトの発展に伴いサポートなどに力を入れるようになる
 - エリート開発者の非技術的な活動への取り組みはバグ修正率との正の相関を除いて, 生産性や品質などのプロジェクトの成果と負の相関がある

所感

- ・ エリート開発者と非エリート開発者の分類方法が参考になった
- ・ 大規模なりポジトリのみの調査となっているため、より小規模なりポジトリで適応できるのかどうか