

Predicting open source contributor turnover from value-related discussions: An analysis of GitHub issues

Jack Jamieson
jack@jackjamieson.net
NTT Communication Science Labs
Kyoto, Japan

Naomi Yamashita
naomiy@acm.org
NTT Communication Science Labs
Kyoto, Japan

Eureka Foong
eureka.foong@gmail.com
Tokyo College, the University of
Tokyo
Tokyo, Japan

Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, Pages 1-13

背景

プロジェクトの価値観

- 社会に良い影響を与えるソフトウェアを構築するのに重要
- 貢献者の参加や離脱に影響を与える可能性がある[25,75]

⇒ 開発中の価値観への関心を高めることが重要

問題点

- 価値観が高まることで、貢献者同士の価値観の対立により、プロジェクトを去ってしまうことが多い

⇒ OSSプロジェクトにおける価値観での議論が必要

既存研究

GitHubのissueでは価値観の議論がされている[37]

問題点

プロジェクトの価値観は時間と共に変化していく

→価値観に対する議論が対立を起こす可能性[18,33]

FilippovaとChoはプロジェクトの価値観に対する対立を調査[19]

- イデオロギー的な意見の相違が議論の有害性の原因[42]
- 価値観に関する議論と貢献者の離脱の関係は自己申告やインタビューに基づいているため、影響の規模を特定するのは困難

目的とリサーチクエスチョン

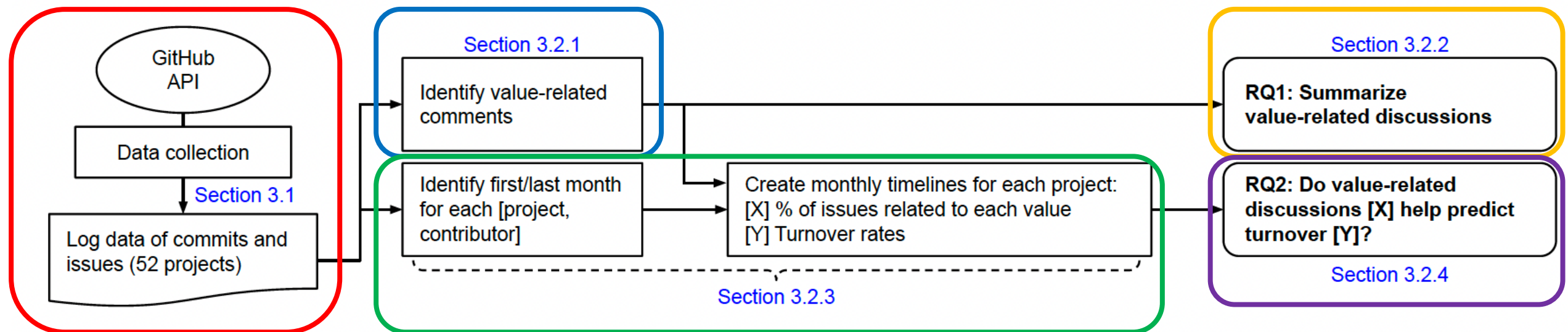
目的

- GitHubのissueで価値観が議論される度合いの変化が、貢献者の出入りに変化をもたらすのかを明らかにすること

リサーチクエスチョン

- RQ1：GitHubのissueにおける価値観に関連する議論にはどのような特徴があるか？
- RQ2：オープンソースプロジェクトにおける価値観に関連する議論の割合は、貢献者の出入りをどの程度予測するか？

データ分析：概要図



- Section3.1 : データ収集
- Section3.21 : 価値観の特定
- Section3.22 : 価値観の要約 (RQ1)
- Section3.23 : 価値観関連のIssueと離職率の定量化
- Section3.24 : 価値観関連の議論と離職率の予測 (RQ2)

データ収集

DWeb (Decentralized Web) に関連するGitHubプロジェクトを対象にする

- Dweb : ウェブの政治経済を再構築するために, 価値観に基づく取り組みを実施^[32,71]

データセット

- リポジトリ数 : 2,179
 - コミット数 : 775,124
 - コメント数 : 783,151
 - issue数 : 277,236
- 「bot」と確認されたものを除いた

価値観の特定・要約（RQ1）

価値観の特定

- ログデータからコメント内に含まれる価値観関連の用語を特定
 - DWebプロジェクトで価値観に関する議論を分析するために使用^[33]
- 本研究で使用する価値観
 - 敬意, 寛容さ, 自由, 公平性と平等, 社会的権力, 環境

価値観の要約

- 各価値観についてランダムに100件の一致を抽出
 1. 価値観の一致が誤検出であるか
 2. 価値観がGitHub上の議論に関連して言及されているか
 3. コメントが価値観に関連した問題を指摘しているか
- } 起こった割合を算出

価値観と離職率の関係について

価値観関連のissueと離職率の定量化

- 価値観関連のissueの定量化

各価値観のIssue

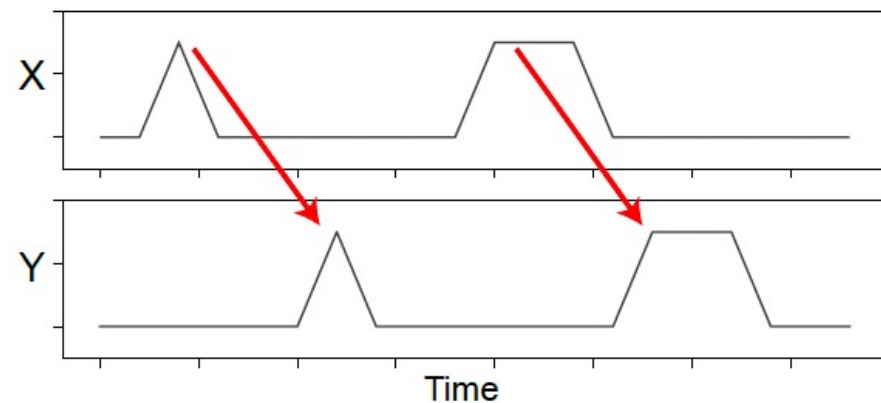
価値観に関する全てのIssue

- プロジェクト離職率の定量化
 - 最初に活動した月「参加月」，最後に活動した月「離職月」
 - Jenks自然ブレイクスアルゴリズム^[34]から平均離職率を4つにグループ分け
- | | |
|----------|-------------------|
| 1：1ヶ月間 | |
| 2：2-3ヶ月間 | 指定した期間内に離職した貢献者の数 |
| 3：4-8ヶ月間 | |
| 4：9ヶ月以上 | 指定した期間に活動した貢献者の数 |

価値観関連の議論と離職率の予測（RQ2）

先行研究^[76]で使用していたGranger分析を使用

- Granger分析の例



⇒ X（議論の割合）の過去値がY（離職率）の現在値に有意な情報を提供するかを測定

- 各価値観と離職タイプの組み合わせについてパネルベクトル自己回帰（pvar）モデルを使用
- Granger分析で有意な結果の場合，インパルス応答関数（IRF）を使用し，さらに評価
 - Xの標準偏差が1増加する際に，Yに与える影響を時系列で測定

RQ1：結果

GitHubのIssueにおける価値観に関する議論にがどのような特徴があるか

Value	issueの 平均割合, 標準偏差		最も頻繁に使われる用語	誤検出	議論を言及	問題の指摘
	Mean	Std. Dev				
Respectfulness	0.19 %	2.33	code of conduct, polite, rude	20.2 %	64.0 %	32.6 %
Freedom	0.21 %	1.26	freedom, user choose, sovereign	27.0 %	0.0 %	11.0 %
Broadmindedness	0.10 %	1.11	diversity, diverse, unconventional	10.4 %	11.5 %	20.7 %
Social Power	0.07 %	1.22	central authority, gatekeeper, monopoly	32.4 %	1.4 %	37.8 %
Equity & Equality	0.03 %	0.27	unfair, fairness, justice	16.7 %	15.3 %	58.9 %
Environment	0.01 %	0.16	climate change, energy consumption, wildlife	15.0 %	0.0 %	17.7 %

特に多かったもの

- 敬意：「議論そのものに価値観を関連付けている」
- 自由：「構築中のソフトウェアを個人が望むように使用する権利を支持」
- 寛容さ：「アイデンティティの多様性や包括性を求める議論」
- 社会的権力：「中央集権的な形態の社会的権力に反対するもの」
- 公平性と平等：「コメントの大部分は不公平, 公平性を含む」
- 環境：「ブロックチェーン技術に関するもの」

RQ2 : 結果

価値に関する議論は離職率を予測するか

帰無仮説：価値観に関するIssueの割合が離職タイプの予測に役立たない

	離職				新規
Turnover type:	Out	Out	Out	Out	In
Contrib. length:	1	2-3	4-8	9+	All
Value					
Respectfulness 敬意	127.701 (0.000) 1 lag	20.4451 (0.000) 1 lag	28.158 (0.000) 1 lag	28.319 (0.000) 2 lags	57.292 (0.000) 1 lag
Freedom 自由	5.018 (1.00) 1 lag	42.404 (0.000) 1 lag	1.588 (1.00) 1 lag	16.865 (0.012) 2 lags	0.411 (1.00) 1 lag
Broadminded- ness 寛容さ	205.009 (.000) 1 lag	0.867 (1.00) 2 lags	4.579 (1.00) 1 lag	—	0.254 (1.00) 2 lags
Social power 社会的権力	2006.876 (.000) 1 lag	2803.359 (.000) 1 lag	144.179 (.000) 1 lag	7.129 (1.00) 2 lags	8.923 (0.672) 2 lags
Equity & equality 公平性と平等	1.422 (1.00) 1 lag	0.015 (1.00) 1 lag	3.761 (1.00) 1 lag	7.739 (1.00) 2 lags	0.174 (1.00) 1 lag
Protecting the Environment 環境	7.081 (1.00) 2 lags	50.097 (0.000) 1 lag	7.634 (1.00) 2 lags	—	15.376 (0.026) 2 lags

57.292
(0.000)
1 lag

Granger因果分析の結果
(p値)

予測に含める過去の月数

全ての変化を予測するのに役立つ

2-3ヶ月間, 9ヶ月以上の離職率を予測できる

長期の離職率を予測するのには役立たない

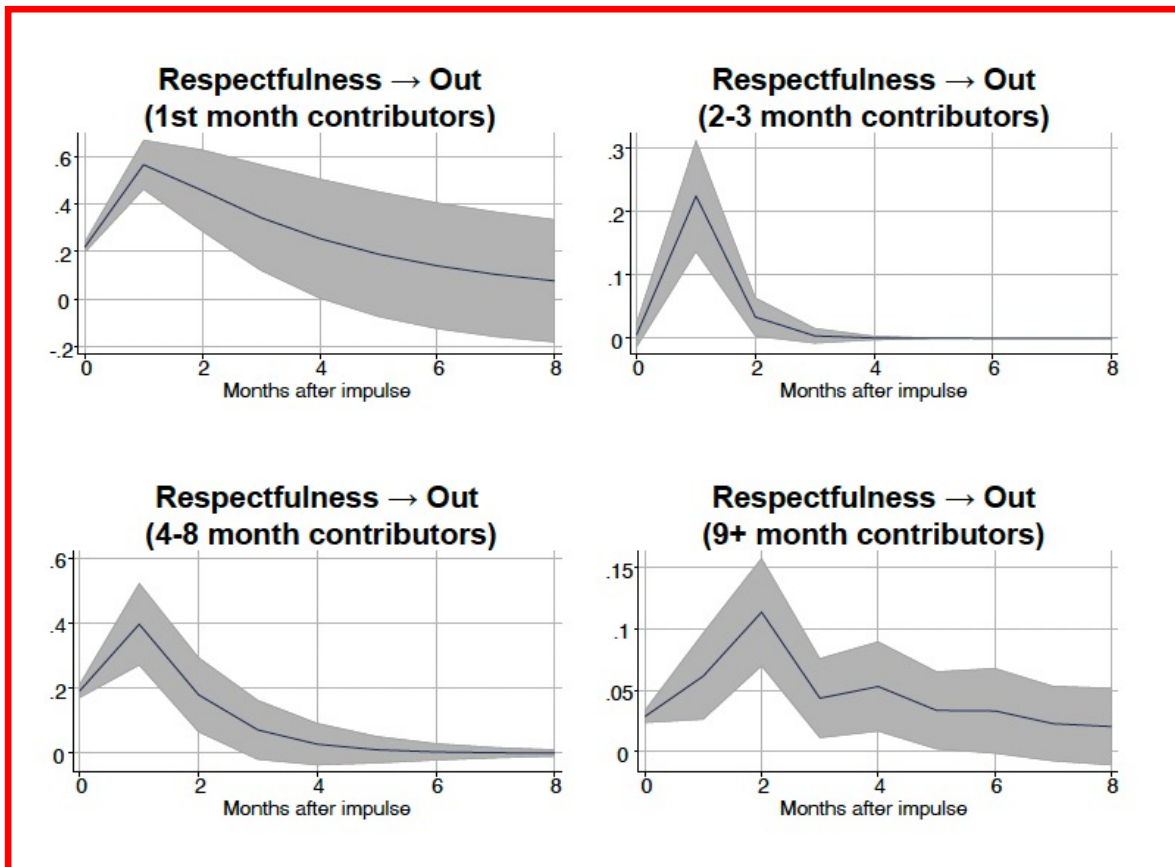
短期間 (1,2-3,4-8) の離職率を予測するのに役立つ

2-3ヶ月間の離職率, 新規貢献者の加入率を
予測するのに役立つ

太字：帰無仮説が5%で棄却されたもの

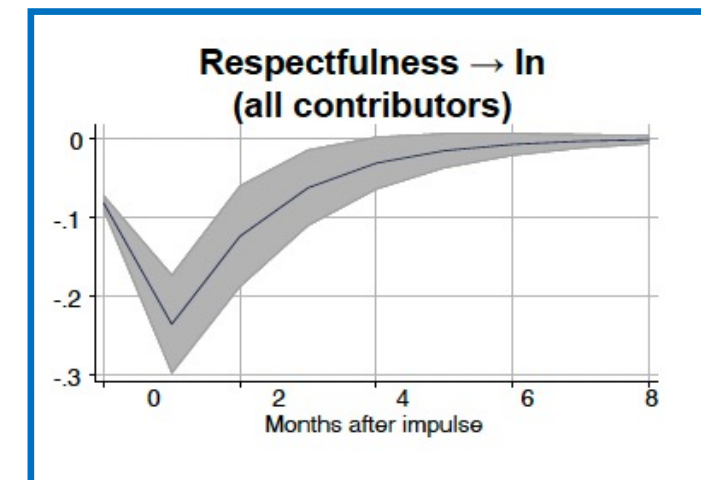
RQ2：結果

離職率変化の予測される規模はどの程度か



離職率のインパルス応答結果

横軸：価値観に関する議論の割合のSD
縦軸：翌月の離職率の変化（効果量）



新規加入率のインパルス応答結果

「敬意」に関する議論の割合が増加すると、
離職率の割合が増加し、同時に新規加入率が少なくなっている

まとめ

目的

- GitHubのissueで価値観が議論される度合いの変化が、貢献者の出入りに変化をもたらすのかを明らかにすること

結果

- 価値観に関するIssueの割合がプロジェクトからの離職率の増加と新規貢献者の減少を予測できることがわかった

補足資料：各価値観の議論例

- 敬意
 - 「他のグループに投稿しながら，ここで応答しないのは失礼だ」
- 自由
 - 「開発者がインスタンスの実装設計をより自由に設計できるようにする」
- 寛容さ
 - 「DAOのジェンダー多様性を高める」
- 社会的権力
 - 「DIDを行うことは，中央権威が関与せずに暗号化マテリアルを見つけることを可能にする」
- 公平性と平等
 - 「新しいアプリの優先が文書化されておらず，不公平と見なされる」
- 環境
 - 「最近，ビットコインの増大するエネルギー消費についての懸念がある」

補足資料：他のインパルス応答

