オープンソースソフトウェアにおける貢献者獲得要因の研究

筑波大学審査学位論文（博士）

2020

小早川 直樹

筑波大学大学院
ビジネス科学研究科 企業科学専攻

概要

今や、オープンソースソフトウェア（OSS）は社会の重要な基盤であり、個人ばかりではなく企業や公共サービスで広く使用されている。そして多くの商用ソフトウェアを駆逐している。OSS の開発にはプログラマーの貢献が欠かせない。プログラマー達はなぜ OSS に引き付けられ、貢献しようとするのか？ OSS の貢献者については、長い間、学術的および実践的な問題として取り組まれてきた。貢献者に関する研究には、プログラマーが OSS に関わる「モチベーション」と「貢献者獲得」のテーマがある。

Linux や Apache プロジェクトが OSS の中心だった頃、開発体制は階層的であり各メンバーの役割と責任が決められていた。メンバー外で可能な貢献は不具合の報告に限られており、より高度な貢献をするためにはメンバーとなる必要があったが、誰でもメンバーになれるわけではなかった。そうした状況下での研究テーマは、プログラマーが OSS に関わる「モチベーション」に関する分析が中心であり、特定のプロジェクトを詳細に調査する研究が多数を占めた。

Linux や Apache 関連のプロジェクトを含む論文数は 2003 年にピークに達し、その後急激に低下した。その後中心になったのがコードホスティングプラットフォーム、特に GitHub である。GitHub は、そのコンセプト「ソーシャルコーディング」に要約されているように、ソーシャル機能に重点を置いている。GitHub は、使いやすい共同ソフトウェア開発ツールと開発者のコミュニティをサポートする多くの機能を提供している。GitHub により、OSS の開発方法が大きく変化した。GitHub では、誰でも「カジュアル」にプロジェクトに参加できる。プロジェクト側の役割もコラボレーター（共同編集者）だけである。GitHub は世界中の多くのソフトウェア開発者の間で人気を博し、多くの OSS が GitHub 上で開発されるようになった。

GitHub のサービスは無償で提供されており、誰でも容易に OSS 開発プロジェクトを始めることができる。しかし、簡単にプロジェクトが始められる状況になると、貢献者の奪い合いが激化する。せっかく斬新なアイデアをもってプロジェクトを開始したとしても、貢献者が集まらず失敗に終わることも多々ある。すなわち、プロジェクト内部の開発者だ

けでなく、外部から開発者（貢献者）を獲得することが成功への鍵となった。そのような背景により、研究テーマも「モチベーション」から、「貢献者獲得」に中心が移っていった。

本研究は、OSS プロジェクトの成功を支援するために、貢献者獲得の要因を明確化することを目的とし、既存研究で可能性を示唆されていた代表的な要因である、インフルエンサー、プロジェクトの将来性、貢献ガイドラインについて、GitHub API などから取得したデータを用い定量的に評価するものである。

はじめに第 2 章において、貢献者に関する先行研究に関して、「モチベーション」と「貢献者獲得」という観点から整理した。「モチベーション」については、多くの研究が E. L. デシ (Edward L. Deci) らによる「自己決定理論」をベースにしており、その用語をそのまま流用している。OSS に貢献する動機は、内発的動機、内在化された外発的動機、外発的動機の 3 つのカテゴリーに分類され、多数の研究がおこなわれた。一方、「貢献者獲得」はライセンスタイプ、スポンサー、インフルエンサーなどが要因として挙げられているが、まだ十分研究されていない。前述したように、誰でも簡単に OSS 開発プロジェクトを始められる昨今においては、「貢献者獲得」は極めて実践的で社会的ニーズが高い研究テーマである。

第 3 章では、貢献者獲得の要因としてインフルエンサーについて分析をおこない、インフルエンサーの存在が貢献者の獲得に有効であることを確認した。また、インフルエンサーの影響力の指標として 3 つの中心性スコア（入次数、PageRank、HITS/Authority）を比較し、妥当性を検証した。分析データには、GitHub 上の仮想通貨プロジェクトから構築したフォローネットワークを用いた。分析の結果、インフルエンサーの影響力と貢献者数との関連性が確認できた。この結果は、影響力のあるユーザーがプロジェクトへの貢献者を集めることに貢献していることを示唆している。また、HITS/Authority スコアが影響力の指標として最も妥当であることが確認できた。

第 4 章では、貢献者獲得の要因としてプロジェクトの将来性について分析をおこない、プロジェクトの将来性が貢献者の獲得に有効であることを確認した。前章と同じく、分析データとして GitHub 上の仮想通貨プロジェクトを使用した。仮想通貨の市場データは公開されており、時価総額の推移が取得できる。時価総額をプロジェクトの将来性を示す代理変数として、貢献者数との関連性を時系列分析手法で分析した。その結果、時価総額が増加（減少）してから 2 か月後に貢献者の数が増加（減少）した。これは、プロジェクトの将来性（すなわち、時価総額）が貢献者をプロジェクトへの参加を促すことを示唆して

いる。

第 5 章では、貢献者獲得の要因として GitHub の標準ファイルである貢献ガイドラインを分析した。構造トピックモデルを用いて記載内容をトピックに分解した後、各トピックと貢献者数との関連性を検証した。その結果、貢献者数と「不具合報告」のトピックの出現確率の間には正の相関、「Git 更新操作」との間には負の相関が存在することが判明した。また、「環境構築」、「コーディング規約」、「Git 操作」、「不具合報告」のトピックの貢献ガイドラインへの記載を提言した。

第 6 章に総括と今後の研究展望について記述する。

本研究において、インフルエンサー、プロジェクトの将来性、貢献ガイドラインが貢献者獲得と関連性があることが確認できた。すなわち、それらの要因に注意を払うことでプロジェクトの成功率を高められる可能性がある。また、HITS アルゴリズムが影響力の指標として効果的であることは今後の研究、貢献ガイドライン記載内容を提言することはガイドラインの質の向上、に寄与するものと期待している。

目次

図目次

表目次

第 1 章

序論

1.1 研究の背景

オープンソースソフトウェア（OSS）は社会の重要な基盤であり、個人ばかりではなく企業や公共サービスで広く使用されている。OSS のソースコードはインターネット上に公開されていて、多数の貢献者により開発やテスト、メンテナンスがおこなわれている。貢献者は、OSS の開発にとって欠かすことができない存在である。現実に、貢献者不足が原因で頻繁に OSS 開発プロジェクトは失敗している [FN08, Kri02, CAH03]。 換言すれば貢献者の数は、プロジェクト成功の指標とみなすことができる。貢献者の数を維持するためには、プロジェクト内部の開発者に加え外部から貢献者を獲得することが重要である。それゆえ、貢献者獲得についての分析が、学術的および実践的な研究テーマとして取り組まれている。

これまで、OSS 開発プロジェクトについて多くの研究がおこなわれてきた。元々の研究対象は、研究者自身が関与するプロジェクトまたは Linux や Apache などのプロジェクトにほぼ限定されていた。当時のプロジェクトの開発体制は、階層的な組織構造であり (図 1.1 左)、各メンバーの役割と責任が決められていた [LC03, MA00, Ray01]。メンバー外で可能な貢献は不具合の報告に限られており、より高度な貢献をするためにはメンバーになる必要があった。しかしながら、一定の試用期間とコアメンバーの公式レビューが必要など、誰でもメンバーになれるわけではなかった。そうした状況下での研究テーマは、プログラマーが OSS に関わる「モチベーション」に関する分析が中心であり、特定のプロジェクトを詳細に調査するものが多数を占めた。ただし、Linux や Apache 関連のプロジェクトを対象とした論文の数は 2003 年にピークに達し、その後急激に低下し

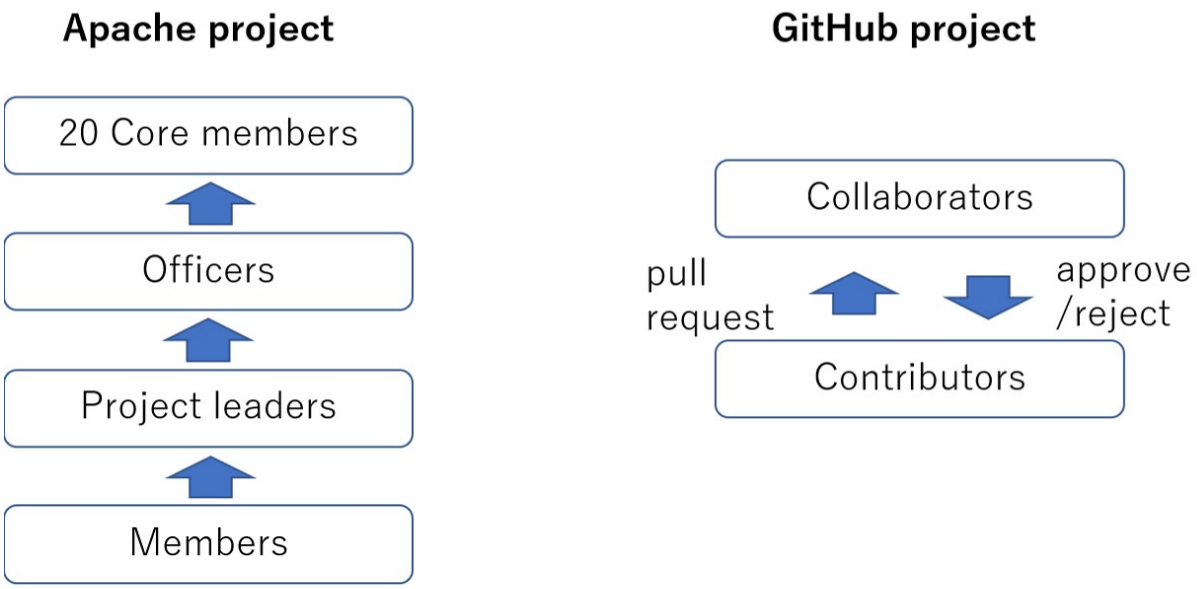
Apache project
20 Core members
Officers
Project leaders
Members
GitHub project
Collaborators
pull
request
approve
/reject
Contributors

図 1.1　開発体制の比較

た [CWHW12]。

次に登場したのが SourceForge、BitBucket のようなコードホスティングプラットフォームである。これらは従来 OSS ごとに個別に用意していたソフトウェア開発環境をインターネット上にサービスとして提供したものである。このサービスにより多数のプロジェクトを一括して分析することが可能となった。その代表が GitHub である。

GitHub の登場 [*1]により、OSS の開発環境が大きく変化した。GitHub は、世界中の多くのソフトウェア開発者の間で最も人気を博している [AA17]。GitHub は、そのコンセプト「ソーシャルコーディング」に要約されているように、ソーシャル機能に重点を置いている。GitHub は、使いやすい共同ソフトウェア開発ツールと、開発者のコミュニティをサポートする多くの機能を提供しており、競合他社と差別化を図っている。

GitHub では、誰でも「カジュアル」にプロジェクトに参加できる。貢献者はプロジェクトのメンバーでなくとも高度な貢献が可能である。プロジェクト側の役割もコラボレーター（共同編集者）だけである。両者のやり取りでソースコードの修正がすすめられる (図 1.1 右)。修正の手順は次のとおりである。

1. 貢献者は貢献したいプロジェクトを「Fork」（プロジェクトの複製を自分の環境に作成）する。
2. 貢献者は「Fork」されたプロジェクトのコードを修正する（新機能の追加、不具合修正）
3. 貢献者は「Pull Request」により、修正したコードを元のコードに統合 (「Merge」) するよう要求する。
4. コラボレーターは、修正されたコードをレビューして、受け入れるか拒否するか決める。
5. 受け入れた場合、修正済コードは元のコードと統合され、貢献は完了する。

GitHub のサービスは無償で提供されており、誰でも容易に OSS 開発プロジェクトを始めることができる。もし魅力的なものであれば貢献者が次々と集まるであろう。このサービスは急速に成長しており、2017 年 4 月の時点で 2,000 万人を超えるユーザーと 5,700 万のプロジェクトが存在する [Git]。 Bootstrap、jQuery、Docker など、今、多くの企業で

[*1] https://github.blog/2008-04-10-we-launched/

使われている OSS は、開発とメンテナンスに GitHub を利用している。

このように誰でも簡単にプロジェクトが始められる状況になると、貢献者の奪い合いが激化する。折角、斬新なアイデアをもってプロジェクトを開始したとしても、貢献者が集まらず失敗に終わることも多々ある。OSS を成功させるには、プロジェクト内部の貢献者だけでなく、外部から貢献者を獲得することが不可欠となり [GLM06]、研究のテーマもモチベーションの理解から貢献者の獲得に中心が移っていった。

GitHub 上には大量のプロジェクトが存在しており、複数のプロジェクトを比較したり、分類したりすることが可能である。GitHub はアプリケーションプログラミングインターフェイス (API) を提供しており、分析アプリケーションから直接各プロジェクトのデータを取得できる。また GitHub の活動データをアーカイブして、そのデータを一括提供する学術的な Web サイト (BOA, GHTorrent など) も存在しており、そのデータも多く利用されている [DNRN13, GS12, GS17]。3 章、4 章では GitHub API、5 章では GitHub API と BOA で取得したデータを利用した。以上の理由により、GitHub に対し極めて多様な研究が行われている [KGB+16]。

研究対象 OSS 開発プロジェクトの変遷を表 1.1 にまとめる。開発環境の変化にともない、関心が移っていったことがわかる。

表 1.1　開発プロジェクトの比較

	2000 年前半	**2000 年代**	**2000 年代後半**
開発環境	個別で用意	コードホスティング プラットフォーム	ソーシャルコーディング
環境例	Linux/Apache	SourceForge	GitHub
開発体制	組織的（右上）	個別	フラット（左下）
貢献者	基本メンバー	個別	誰でも（カジュアル）
分析プロジェクト	単独〜数件	多数（数十万）	多数（数千万）
情報公開	Web で公開	API	API、アーカイブなど
関心	モチベーション の理解	個別	プロジェクトへの 貢献者獲得

次章において詳細なレビューをおこなうが、OSS 開発の貢献者に関する先行研究は、

「モチベーション」と「貢献者獲得」という分野の違い以外に、「定性的分析」と「定量的分析」という手法の観点、「単独プロジェクト分析」と「プロジェクト集合分析」というデータの観点から整理できる。時間的には前者から後者に研究の中心が移っている。

「モチベーション」とは OSS に貢献する動機を特定する研究であり、「貢献者獲得」とはプロジェクトが選択された要因を特定する研究である。「定性的分析」とはプロジェクト参加者にインタビューやサーベイをおこないその結果を分析し知見を得る手法、「定量的分析」とはプロジェクトの活動データを収集して分析し解釈する方法である。近年では、その両方を組み合わせた研究も多々見られる。「単独プロジェクト分析」とは、Apache など特定のプロジェクトについて組織論などを踏まえ詳細な分析をおこなう方法である。一方「プロジェクト集合分析」は GitHub などから多数のプロジェクトの情報を取得し一括して分析する方法である。先行研究の分類についてまとめたものを、図 1.2 に示す 。

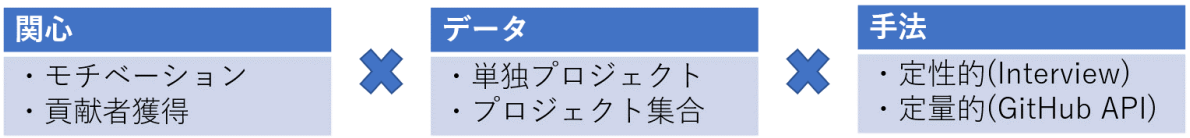

図 1.2　先行研究の分類

1.2　本研究の位置づけ

本研究の目的は、OSS 開発プロジェクトの成功を支援するために、貢献者獲得の要因を明確化することである。 先行研究を踏まえた本研究の位置づけは、貢献者獲得の要因について「プロジェクト集合」から取得したデータを用いた「定量的分析」となる。先行研究として、モチベーションの分析は多数存在するものの、貢献者獲得についてはまだその要因が、十分研究されていない。特に、インフルエンサー、プロジェクトの将来性、貢献ガイドラインについては、以下で説明するように課題が多く残されており、本研究におけるテーマとして選択した。ところで、貢献者の定義であるが、本研究では先行研究同様、プロジェクトの成果物に対し一度でもコミット (変更) したユーザーとしている [BNM+11]。

- 近年のソーシャルメディアの研究ではインフルエンサーの存在が注目を集めてい

る [CHB+10, BS16]。インフルエンサーは専門分野（ドメイン）の伝道者であり、他の人を説得し誘導する重要な役割を果たすとみなされている [DG14]。インフルエンサーは貢献者の行動に影響を与えると考えられており、OSS 開発における重要な研究対象である。しかしながら、従来の研究は GitHub 全体のインフルエンサーの分析に留まり、特定のドメインでの分析は未実施であった。また、従来の研究には問題があった。1 つは解析対象のプロジェクトの抽出である。インフルエンサーの特定は基本的にネットワーク分析でおこなう。つまり、ネットワーク構造が維持されていないと信頼性が低下する。しかし、先行研究においてはネットワーク構造維持に配慮したプロジェクト抽出がおこなわれていなかった [TBLJ13, CLC16]。もう 1 つは、ネットワーク分析の手法である。研究ごとに異なった手法が用いられており、どの手法が有効か検証されてこなかった。これらの課題を解決した上で、インフルエンサーの影響を調査する必要がある。

- 将来性のあるプロジェクトに貢献することは、ユーザーの評判やキャリアにとって有利に働くであろう、一方、プロジェクトに将来性がなければ、彼らの貢献はすべて無駄になってしまうかもしれない。Dabbish らはインタビューによる調査の結果、ユーザーがどのプロジェクトが長期的に成功するか、公開されている情報を基に判断していることを発見した [DSTH12]。プロジェクトの将来性は、貢献者獲得につながる重要な要因と推測される。しかし、実際にプロジェクトの将来性が貢献者獲得に寄与してるかについては検証されてこなかった。将来性を示す代理変数を設定し、将来性と貢献者数との関連性を明確にしていく必要がある。
- ソフトウェア開発においてガイドラインは根幹をなす。ガイドラインが不備であると、ユーザーに不信感を与えるだけでなく、規則性が失われ不具合を生じやすくなり、メンテナンスの負荷が増大する。GitHub では貢献者のために、貢献ガイドラインという標準ファイルを設けている。貢献ガイドラインの効果を検証し品質を向上させることは、ユーザーに信頼感を与え貢献者獲得に寄与するだけでなく、プロジェクトのメンテナンスの負荷を軽減させることにつながる。

上記の 3 つの要因は以上の様に研究対象として十分意義があり、研究テーマとしてふさわしいと考えられる。

1.3　本研究の構成

本論文の構成は次のとおりである。

２章においては、まず、貢献者のモチベーションの研究について整理する。貢献者の研究は、基本的にモチベーションの研究を出発点にしており、正しく理解しておく必要がある。近年はモチベーションから貢献者獲得に研究の中心が移っており、本研究でも貢献者獲得に特に焦点をあてて研究する。次に現状の貢献者獲得の研究について整理する。最後に３章以降の研究に関する先行研究についてまとめる。

次の３章では、インフルエンサーについてその存在が貢献者の獲得にどのような影響を及ぼすか検証した。インフルエンサーの分析は、Twitter などソーシャルメディアでは人気が高いが、GitHub ではまだ少ない。また、インフルエンサーの特定はネットワーク分析でおこなわれるのが一般的だが、様々な手法が存在しどの分析手法が有効か検証されていない。本研究では、仮想通貨プロジェクトからフォローネットワークを構築し、それに対し入次中心性、PageRank、HITS アルゴリズムを用いインフルエンサーを特定し比較した。そしてその最良アルゴリズムで得た指標を用い、インフルエンサーが貢献者獲得に効果があるか検証した。

４章では、仮想通貨の時価総額を将来性の代理変数として、貢献者数に与える影響を分析した。仮想通貨を構成する主要なソフトウェアは GitHub 上で開発、メンテナンスされている。それにより仮想通貨プロジェクトの属性を GitHub API で取得できる。また仮想通貨の時価総額も一般提供されており、仮想通貨プロジェクトの各属性と時価総額を組み合わせた分析が可能である。仮想通貨の時価総額の変化が仮想通貨プロジェクトの貢献者数に及ぼす影響について、時系列分析手法を用い検証した。

５章では、貢献ガイドラインと貢献者数との関連性について検証した。貢献ガイドラインは GitHub の標準ファイルであり、開発者が特定のプロジェクトに貢献する際に参照する為に設けられている。これまで貢献ガイドラインは十分に研究されてこなかった。特に、その記載内容まで踏み込んだ研究はほとんど存在しない。ソフトウェア開発においてガイドラインは根幹をなす。しかし、貢献ガイドラインの記載内容は定義されておらず、ユーザーを混乱させている。本研究では、構造トピックモデル (Structural Topic Model) を用い、貢献ガイドラインをトピックに分解した後、各トピックと貢献者数との関連性を

分析した。また、貢献ガイドラインへの記載内容を提言した。

6章に総括と今後の研究展望について記述する。

第 2 章

OSS の貢献者に関する先行研究

貢献者数を確保する為には、貢献者について理解することが必要である。前述したように貢献者に関する研究には、「モチベーション」と「貢献者獲得」がある。両者に関する先行研究を俯瞰する。また、3 章以降の研究に関する先行研究についても、まとめて説明する。

2.1　モチベーションに関する先行研究

プログラマー達が自分でプロジェクトを完了するために必要なすべての作業をおこなわなければならず、しかもプロジェクトがほとんど宣伝されないにもかかわらず、なぜ彼らはソースコードを公開し、ソフトウェアを無料で提供することを選ぶのか？ プログラマーが OSS へ関わるモチベーションについては、長年にわたって研究されてきた。そこで使われている理論は、次に述べる「自己決定理論」をベースとしており、用語も流用されている。

2.1.1　自己決定理論 (Self-Determination Theory)

E. L. デシ (Edward L. Deci) らによる「自己決定理論」(Self-Determination Theory) [DR80] では、モチベーションの高低を、量的な差異ではなく、質的な差異によって分かれる六つの段階として示している。これら六段階を、モチベーションの低いものから順に示すと、次のようになる。 [DF95, RD00, 中谷 07]。

- 非動機 (Amotivation) づけ。ある行為を、たまに何かの動機があってやりはして

も、やり続けることは無い状態。

- 外発的 (Extrinsic) 動機づけ、外的調整 (External regulation) 。ある行為を、報酬を得、懲罰を避けるために、やり続けている状態。
- 外発的動機づけ、取り入れ的調整。ある行為を、名誉心や恥ずかしさから、やり続けている状態。
- 外発的動機づけ、同一化的調整。ある行為を、目標や成長に必要だから、やり続けている状態。
- 外発的動機づけ、統合的調整。ある行為を、やるのが自然なこととなって、やり続けている状態。
- 内発的 (Intrinsic) 動機づけ。ある行為を、それをやること自体が楽しいから、喜んでやり続けている状態

2.1.2 動機を３つのカテゴリーに分類

Krogh らは過去の研究を調査し、OSS に貢献する動機を「自己決定理論」に基づき、内発的 (Intrinsic)、内在化された外発的 (Internalized Extrinsic Motivation)、および外発的 (Extrinsic) の 3 つのカテゴリーに分類した [KHSW12]。内発的な動機には、イデオロギー、利他主義、友人関係、趣味、娯楽などが含まれる。内在化された外発的動機には、評判、互恵関係、学習、利用価値などがあり、外発的動機には、キャリアと報酬が含まれる。それぞれのカテゴリーに関する研究を以下に紹介する。

内発的動機

内発的動機について、最初に頭に浮かぶのはストールマンの「ソフトウェアはすべて無料にすべきであり、自由に修正できて配布されるべきだ」というイデオロギー [Sta06] への信奉だろう。1588 名の OSS 開発者を調査したところ約 8 割がこの考え方に賛同していた [DWA03]。それ以外にも、社会的意義への意識が高いことと開発生産量が関連している [HNH03]、内発的動機が高い人は業務により集中しリスクを取りたいと考え、常に代替戦略を模索している [Hen00]、など肯定的な研究が多い。ただし、実際にプロジェクトに関わった理由はばらばらである [DWA03]、利他的およびコミュニティへの帰属意識などは重要な動機ではあるが、プロジェクトへの参加決定は、報酬など外発的な動機が重要で

ある [OF00, AH02] など、内発的動機は貢献に直接結びつかないとの指摘がある。

内在化された外発的動機

Ghosh らは 2700 人の開発者にサーベイをおこない、OSS 開発プロジェクトに参加し今も留まっている最も重要な理由は、学習とスキルの向上であるとした。利他的な動機は見いだせなかった [Gho05]。 Lakhani らは Apache Project の技術サポートという単調な仕事が、互恵関係で動機づけられているとした。過去に誰かに助けられた担当者は、次は自分が助けなければならないとの意識が高い傾向にあった [LVH04]。Lattemann らは、動機は立場によって異なるとして、動機を体系化した。（不具合修正者や管理者よりも）プログラマーは周りの評判によって動機づけられるとした [LSR08]。不具合修正者は自分で利用するためにプロジェクトに参加する [VH01] などの研究もあり、立場や条件によって異なった動機が示されている。

外発的動機

外発的動機が直接的に貢献につながる、とする研究は多数存在する。以下にその研究例を示す。OSS の経済は、開発者のキャリアの懸念によって説明できる。開発者は自分の開発したソフトウェアが公開されることで、自らの才能を潜在的な雇用主に知らせ、労働市場での価値を高められると考えている [LT03]。報酬を受けている貢献者は週当たり 17.7 時間費やすのに対し、無報酬の貢献者は 11.7 時間しか費やさなかった [LW03] 。Linux Kernel の開発の半分以上は、仕事として携わっている開発者によるものである [K$^+$07] 。インセンティブがパフォーマンスに直接関連している場合、内発的な動機はパフォーマンスにとってそれほど重要ではない [CNF14]。内発的動機が高くないユーザーにとって、報酬は貢献量に直接プラスの影響を及ぼす [AL11]。Robert らは、動機と貢献量との関係を研究した。彼らは、Apache Software Foundation の傘下にある 3 つの主要な OSS プロジェクトを分析し、開発者の報酬とステータスに関連する動機は直接貢献量に寄与するが、内発的動機は直接的には貢献には結びつかないとした [RHS06]。

まとめ

以上を総括すると、実際にプロジェクトに関わるという決断には、外発的動機、たとえば報酬とかキャリアなどが重要であると考えられる。他の動機についてはユーザーの立場や条件によって様々である。内発的動機は肯定的ではあるが、貢献への直接的な効果は

薄いと思われる。ただし、創造的な作業については報酬は機能しない [BN12, AGLM09] など、外発的動機があてはまらないケースの指摘もある。また、当初の動機と継続的な貢献は無関係 [HK03, FN08]、ステータスへの動機が娯楽性を高める [RHS06]、多くのユーザーが一度限りの貢献ではあるが、一部のユーザーは趣味、娯楽的な動機に内発されて、そのまま長期的に貢献活動を続ける [Sha06]、など当初は外発的動機で参加したものの、継続して貢献していくうちに内在化していくことが観察されている。Krogh らは、そのことについて「目の前のニンジンへの追求は、虹のかなたにある目に見えない大きな報酬によって置き換えられる」 [KHSW12] という言葉でまとめている。

2.2 貢献者獲得に関する先行研究

2.2.1 時代背景

OSS 開発プロジェクトに関する研究が始まった当時、対象となるプロジェクトは限定されていた。研究対象は、研究者自身が関与するプロジェクト [NYN+02, YK03] か、Linux、Apache プロジェクトであった。Linux は明らかに最も一般的に研究された OSS 開発プロジェクトであり、次に Apache（通常は httpd サーバーを意味する）が続く。ただし、これら 2 つのプロジェクトを対象とした論文の数は 2003 年にピークに達し、その後急激に低下した [CWHW12]。その後、Wikipedia [OO07, ON08] なども対象となり、対象となるプロジェクトも増加したが、基本的に単独プロジェクトの分析が中心であった。

2000 年代になり、SouceForge や BitBucket などのコードホスティングプラットフォームが登場。その後 GitHub が人気となり、開発環境が大きく変化した。研究テーマもそれに伴い変わっていった。今までは、モチベーションの研究が中心であったが、GitHub の登場によりプロジェクト間の競争が始まると、モチベーションの理解だけでは不十分となり貢献者獲得の要因に関する研究に中心が移っていった [PSG16, PSL+13]。また GitHub API で取得したデータを用いた定量的な分析による研究が増加した [ICC15, GPD14]。GitHub 上には多数のプロジェクトが存在するため、プロジェクトの集合が分析されるようになり、プロジェクト間の比較や分類に関する研究がおこなわれるようになった [YMKU14, YKM+16]。

総括すると、近年はモチベーションから貢献者獲得に研究の中心が移っている。本研究でも貢献者獲得に特に焦点をあてて研究する。

2.2.2 貢献者獲得の先行研究

Stewart らは社会スポンサーシップの役割を調査し、営利（企業など）のスポンサーと非営利（大学など）のスポンサーを区別した。そして、プロジェクトのスポンサーのタイプが開発者の貢献へのモチベーションに影響を与えていると結論づけた [SAM06] 。たとえば、非営利スポンサーのプロジェクトは、スポンサーをもたないプロジェクトよりも多くの開発活動量を得られることを確認した。ただし、GitHub が登場する前の研究で取得できる属性に制約があったため、リリース頻度を開発活動量の代理変数としており、本当に個人が活動量を増やしているのか検討の余地がある。Dabbish らは、開発物などユーザーが GitHub で公開された情報を基に、長期的にどのプロジェクトが繁栄するか推測していることを発見した [DSTH12]。また、Tsay らは同じデータを用い、ユーザーはプロジェクトを選択する際、プロジェクトの貢献者のステータスを参考にしているとした [TDH13]。Dabbish と Tsay が用いたデータは、公開された情報からユーザーが推測していることをインタビューを通じてまとめたものであり、推測した結果の行動まで踏み込んでいない。それ以外にも、制限の少ないライセンスタイプほど貢献量は増加する [FG07] 、頻繁なコードのリリースや文書の改善は、貢献者を引き付ける [BHV16, AHS14] などが観察されている。

序論で述べたように、貢献者獲得の要因については課題が多く残されている。本研究では、インフルエンサー、プロジェクトの将来性、貢献ガイドライン 3 つの要因を分析した。各分析においては、それぞれ、ネットワーク分析、時系列分析、テキストマイニングの手法を用いた。各要因と各手法に関連した先行研究を以下に説明する。

ネットワーク分析とインフルエンサーの先行研究

ネットワーク分析手法を用いて GitHub のデータを分析した研究は以下が存在する。Thung らは、共通の開発者で紐づけられたプロジェクトネットワークと、共通のプロジェクトで紐づけられた開発者ネットワークをそれぞれ構築し、影響力のあるプロジェクトと開発者を抽出し分析した [TBLJ13]。彼らは、GitHub API を使用して取得した最初の 100,000 プロジェクトをデータとして使用した。Yu らは GitHub のアーカイブデータサイトである GHTorrent [*1]から抽出した約 180 万人のユーザーのデータからフォローネッ

[*1] https://ghtorrent.org/

トワークを作成し、パターンを抽出したうえでタイプ分類した [YYWW14]。いずれの研究でもネットワーク構造の維持に配慮せずにデータを抽出しており、ネットワークが分断された可能性がある。3 章に示した研究では、仮想通貨プロジェクトを採用したことにより、その問題を回避した。

GitHub 上のプロジェクトのインフルエンサーに関する主要な研究は主に以下の 3 つに分類される。

1. インフルエンサーの影響力の指標は何か？

   Thung らは PageRank アルゴリズムを用い、そのスコアをインフルエンサーの影響力の指標とした [TBLJ13]。Blincoe らはインフルエンサーの影響力の指標としてフォロワー数を用いた [BSG+16]。 GitHub にはさまざまなコラボレーション機能が存在する。「フォロー」はユーザー間のリンク機能であるが、「Fork」（ソースコードのコピー）、「Watch」（プロジェクトをフォロー）、「Star」（プロジェクトをブックマーク）は、ユーザーとプロジェクト間をリンクする機能である。Badashian らは、（ユーザーが所有する）プロジェクトの「Fork」と「Watch」数を、フォロワー数とともに考慮する必要があると主張した [BS16]。しかしながら、どの指標が適切か検証した研究は存在しない。
2. インフルエンサーになるためには多大な貢献活動が必要なのか？

   従来の研究では、インフルエンサーになるために多大な貢献活動は必須でないとしている。 GitHub で多くの活動をおこなうユーザーは、必ずしもインフルエンサーとは限らない [LRM14, BSG+16]、GitHub と同じユーザー名を使用してブログや Twitter で自分自身を宣伝するなどの活動が、影響力を持つために必要である [DSTH12]、など否定的な研究が多い。対照的に、Twitter のインフルエンサーになるには、頻繁にツイートするなどの相当な努力が必要である [CHB+10]、など他のソーシャルメディアでは肯定的に結論づけられている。GitHub には様々なプロジェクトが存在しており、一律に当てはまるのか検証が必要と考えられる。
3. インフルエンサーの存在により、貢献者をより多く獲得できるか？

   先行研究において、インフルエンサーの存在は貢献者獲得に効果があるとしている。ユーザーはプロジェクトを選択する際、プロジェクトの貢献者のステータスを参考にしている [TDH13]、インフルエンサーはフォロワーを新しいプロジェクトに

導く [BSG+16, KY19] などの研究が存在する。

本研究（3 章）では上記 3 つのテーマを、仮想通貨プロジェクトに適用し、仮想通貨ドメインのインフルエンサーの影響力を分析した [KY19]。

時系列分析と将来性の先行研究

ユーザーが GitHub で公開された情報を基に、プロジェクトの将来性を推測しているとした研究 [DSTH12] は存在するが、推測した結果、実際に貢献に結びついたかまで分析した研究は存在しない。また将来性の指標を何にするのかが課題である。時間的な変化を考慮した研究は、数は少ないがいくつか存在する [Sha06, SAM06]。たとえば、プロジェクトは時間の経過とともに一時的な貢献者の割合が低下して、コアメンバー中心の体制に移っていく [GM11] などが観察されている。因果関係を定量的に分析するには、時間差を利用した調査が実用的である [Gra69]。本研究（4 章）では時系列分析により、仮想通貨の時価総額を将来性の代理変数として、貢献者数との因果関係を分析した [KINY20]。

貢献ガイドライン分析の先行研究

GitHub は、貢献ガイドラインと呼ばれる標準ファイル（contributing.md/contributing.txt [Bar]）を各プロジェクトで用意することを推奨している。Izquierdo らは、貢献量（コミット数）の多い上位 50 プロジェクトの調査をおこない、そのうち 46 プロジェクトが貢献内容について記述をしたガイドラインもしくは Web サイトを用意していたと報告している [ICC15]。さらに、その一部を調査し、貢献についての説明と管理規則を明確に記述することが、新しい貢献者を引き付ける重要な要素であるとした。ただし予備的な調査 (数例の主観的な分析）であり、また貢献者を引き付ける項目が何であるか特定していない。Chen らは、貢献ガイドラインを含むプロジェクトは、含まないプロジェクトよりも 25%〜45% 生存する可能性が高いことを示した [CP16]（7 日以内のコミットを生存基準とした場合、7.2% に対し 5.0%、30 日間の場合 14.8% に対し 11.8% ）。ただし、ガイドラインの存在に関する調査であり、ガイドラインの記載内容まで踏み込んで分析していない。

本研究（5 章）は、貢献ガイドラインの記載内容まで踏み込んだ研究 [KY17] を、より詳細に説明するとともに、構造トピックモデルを用い解析方法を改善し拡張したものである。前研究の発表後、それをリファレンスとして、複数の研究が発表されている [DBR+18, WMWS19, ESEZ19]。たとえば、Elazhary らは 53 の貢献ガイドラインの

記載内容を調べ、実際の活動がそれに従っているか分析した。そして開発プロジェクトの 68% がガイドラインと乖離した活動をしていることを明らかにした [ESEZ19]。

貢献者獲得に関連した主な先行研究を表 2.1 にまとめた。

表 2.1　貢献者獲得関連の先行研究

	要因	先行研究	データ	分析手法	備考（課題）
プロジェクトの形態	オーナー	Stewart2006 [SAM06]	138 OSS projects from www.freshmeat.net	回帰分析	スポンサーあり。(情報が少ない (リリース頻度、ML 登録とか）)
	ライセンスタイプ	fershtman2007 [FG07]	71 most active projects hosted on SourceForge	回帰分析	制限が厳しくないもの (サンプル少ない)
体制	インフルエンサー	Blincoe2016 [BSG+16]	199 popular (most followed) users and their followers on GitHub archive（GHTorrent)	Survey & 回帰分析	(プロジェクト抽出の根拠が薄いなど)
	参加者のステータス（評判）	Tsay2013 [TDH13]	24 GitHub users	Interview	(実証分析が必要)
成果物	将来性（開発物などから）	Dabbish2012 [DSTH12]	24 GitHub users	Interview	(具体性はない）(実証分析必要)
	頻繁なリリース	borges2016 [BHV16]	2,500 popular projects in GitHub	記述式統計	新機能のリリース頻度（プロジェクト抽出の根拠）
	頻繁なリリース	aggarwal2014 [AHS14]	90 projects in GitHub (MSR’ 14 event)	回帰分析	文書のリリース頻度
	貢献ガイドラインの記述	izquierdo2015 [ICC15]	7,365,622 projects from GitHub archive(GH Archive)	相関分析	（予備的な研究）（本格的な分析が必要）
	貢献ガイドラインの存在	chen2016 [CP16]	70,000 projects from GitHub archive (BOA)	記述式統計	貢献ガイドライン数は 170 (記載内容まで踏み込んでいない)

2.3 貢献者獲得の先行研究の課題と本研究での対応

インフルエンサーは専門分野（ドメイン）の伝道者であり、他の人を説得し誘導する重要な役割を果たすとみなされている。 しかし GitHub における先行研究では、ドメインではなく GitHub 全体でのインフルエンサーの分析に留まっていた。3 章では、仮想通貨プロジェクト集合を用い、ドメインのインフルエンサーを分析した。また、インフルエンサーの影響力の指標が各研究で異なっており、研究ごとの比較が困難、指標が不適切、などのリスクがある。GitHub での影響力の指標については未検証であり、喫緊の課題である。3 章では、3 つの著名な指標（入次数、PageRank、HITS/Authority）を比較し妥当性を確認した。

将来性のあるプロジェクトに貢献することは、ユーザーの評判やキャリアにとって有利に働くであろう、一方、プロジェクトに将来性がなければ、彼らの貢献はすべて無駄になってしまうかもしれない。Dabbish らは、ユーザーが公開された情報から、どのプロジェクトが将来的に繁栄するのか推測しているとした [DSTH12]。しかし、推測した結果、実際に貢献に結びついたかまでは踏み込んでいない。それを実証する研究が必要である。ただし、将来性の指標が課題であった。仮想通貨には、市場価格や取引量（ボリューム）などの金融情報が存在する。4 章では、仮想通貨の時価総額を取得し、それを将来性の代理変数とし、時系列分析を用い検証をおこなった。

貢献ガイドラインが不備であると、規則性が損なわれメンテナンスの負荷が大きくなる。実際、貢献ガイドラインは記述レベルの差が顕著であり、存在だけの研究 [CP16] では不十分であり記載内容まで発展させる必要があった。記載内容については、主観に基づいた予備的な分析に留まっていた [ICC15]。5 章では、構造トピックモデルを用い、貢献ガイドラインの記載内容まで踏み込み客観的に分析した。

研究全般の課題として、プロジェクトの抽出方法がある。GitHub 上のプロジェクト数は巨大であり、一括で分析することは不可能である。したがって、一部の抽出が必要とされるが、最初の 10 万プロジェクト [TBLJ13]、人気のあるプロジェクト [BSG+16, BHV16] など根拠のない抽出が頻繁に行われている [CLC16]。特に、インフルエンサーに関しては、ネットワーク構造を維持しなければならないため、注意して抽出する必要がある。本研究では、3 章と 4 章の研究で仮想通貨という共通のドメインに所属するプロジェクト、

5 章で貢献ガイドラインを含むプロジェクト（貢献者を必要としてるプロジェクト）、など明確な方針をもって抽出をおこなっている。仮想通貨ドメインは 3 章で詳しく説明するが、仮想通貨プロジェクトの一覧が公開されていることで客観的なプロジェクトの抽出が可能である。

最後に、貢献者獲得の先行研究の課題と本研究での対応をまとめたものを表 2.2 に示す。

表 2.2　貢献者獲得の先行研究の課題と本研究での対応

貢献者獲得要因	先行研究とその貢献	課題（リスク）	本研究での対応
インフルエンサー	インフルエンサーが貢献者を獲得（GitHub 全体） [BSG+16]	ドメインインフルエンサーの方が効果的だが未検証	仮想通貨ドメインのインフルエンサーを分析
	PageRank などを影響力の指標として分析 [TBLJ13, BSG+16]	影響力の指標は未検証 不適切な指標であるリスク	指標を比較して最適な指標で分析
	ネットワーク分析によりインフルエンサーを特定 [TBLJ13, YYWW14]	プロジェクトをランダムに抽出 ネットワーク分断のリスク	ネットワーク構造を維持するように抽出
プロジェクトの将来性	公開情報から将来性を判断できる事を指摘 [DSTH12]	判断後の行動は研究の範囲外	貢献への寄与を確認
	将来性について言及 [DSTH12]	将来性は抽象的な表現に留まる	仮想通貨の時価総額を将来性の代理変数とする
貢献ガイドライン	ガイドラインの存在が生存確率を高めることを指摘 [CP16]	記述レベルに差がある	ガイドライン記載内容を分析
	記載内容が重要であると指摘 [ICC15]	主観に基づく予備的な分析 実証分析が必要	構造トピックモデルによる客観的な分析
全般	GitHub からプロジェクト抽出（先頭の n 件や人気のある）	根拠のない抽出が存在 結果が偏るリスク [CLC16]	仮想通貨プロジェクト、貢献ガイドラインのあるプロジェクト

第 3 章

ネットワーク分析を用いたインフルエンサーと貢献量との関連性の研究

本章では、貢献者獲得の要因としてインフルエンサーについて分析をおこない、インフルエンサーの存在が貢献者の獲得に有効であることを確認する。また、インフルエンサーの特定方法について複数のアルゴリズムを比較し、妥当性を検証した。具体的には、GitHub 上の仮想通貨プロジェクトのフォローネットワークを構築し、ネットワーク分析手法を用い、分析をおこなった。

3.1 序論

近年のソーシャルメディアの研究ではインフルエンサーの存在が注目を集めている [CHB+10, BS16]。インフルエンサーは専門分野（ドメイン）の伝道者であり、他の人を説得し誘導する重要な役割を果たすとみなされている [DG14]。 Twitter などのソーシャルネットワーキングサービス（SNS）でインフルエンサーのさまざまな分析がおこなわれている。GitHub などのコードホスティングプラットフォームにおいても、インフルエンサーの研究も始まった。インフルエンサーは貢献者の行動に影響を与えると考えられているが、インフルエンサーを特定する方法、インフルエンサーが貢献者に与える影響については十分に研究されていない。特に、特定のドメインにおけるインフルエンサーの研究は存在しない。ドメインインフルエンサーは、特定のドメインに関連付けられていないインフルエンサーよりも、特定ドメインのユーザーに効果的なアプローチが期待できる。

本章では、仮想通貨のドメインインフルエンサーに着目した。現在、多くの仮想通貨が

OSS として開発されている。最初の仮想通貨であるビットコインも OSS である。ビットコインは、Satoshi Nakamoto という名称の個人またはグループによって発明された。その匿名の開発者は仮想通貨に関する論文を 2008 年 [Nak08] に公開し、それをもとにしたビットコインを 2009 年にリリースした。興味深いのは、誰もがソースコードを閲覧できるだけでなく、ソースコードを分岐して独自の仮想通貨プロジェクトの立ち上げができることである。ビットコイン以後、いくつかの異なるスキームの仮想通貨が考案され、そのほとんどが GitHub 上で開発または維持されてきた。現在、GitHub 上には 600 を超える仮想通貨プロジェクトが存在し、すべての主要通貨が含まれている（次章の表 4.1)。それらのプロジェクトの集合および関連するユーザーを、ここでは仮想通貨ドメインと定義する。

仮想通貨ドメインの研究にはいくつかの利点がある。まず、仮想通貨プロジェクトの一覧が公開されていることである[*1]。一般的に、ドメインに所属するプロジェクトを選択するのは主観的になりがちである。この一覧により、仮想通貨ドメインに所属するプロジェクトを客観的に抽出できる。次に、仮想通貨の注目度が高いことと、そのオープン志向である。それらのおかげで、金融情報サイト、ブログ、ソーシャルメディアなどから、豊富な情報を取得できる。

仮想通貨ドメインからインフルエンサーを特定し、影響を測定するためのデータとして、仮想通貨プロジェクトのユーザー間をフォロー関係で結び、フォローネットワークを作成した。「フォロー」は GitHub の標準機能である。あるユーザーをフォローすると、そのユーザーが GitHub 上で活動する度に、イベントとして通知が送信されてくる。このフォロー関係を有向グラフとしたものが、フォローネットワークである (図 3.1)。フォローネットワークの分析により、ドメインの構造的特徴を定量的に明らかにすることができる。詳細は 3.3 章に記述する。

プロジェクトの抽出は、多数のプロジェクトを含む GitHub では重要事項である。 2017 年 4 月の時点で、GitHub は 5,700 万のプロジェクトと、2,000 万を超えるユーザーを抱えている [Git]。全データを分析することは現実的ではないため、先行研究では、人気のあるプロジェクトに限定するなど一部のデータを抽出して分析に用いている。ただし、抽出方法によっては偏った結果を生み出すリスクがあると指摘されている [CLC16]。特にノー

[*1] http://coinmarketcap.com（accessed 2018-02-12)

ド（ユーザー）間の関連性を分析するネットワーク分析においては、ネットワーク構造を可能な限り損なわないようにデータを抽出する必要がある。ランダムにノードを抽出すると、始点ノードまたは終点ノードの一方しか抽出されない可能性が高く、ノード間のリンクが失われネットワークが分断されてしまう。本研究でのデータセットは、同じドメインに所属するプロジェクトのユーザーであるため、フォローするユーザーとフォローされるユーザーの両方が含まれている可能性が高い。したがって、ネットワークの分断は少ないと推測される (図 3.2)。

本章では、仮想通貨ドメインでの影響力を測定するための指標を検証し、その後、インフルエンサーの影響力と貢献との関連性を分析した。リサーチクエスチョンは次のとおりである。

RQ 1： 仮想通貨インフルエンサーをどのように特定できるか？

2 つの異なるタイプのインフルエンサーが存在する。仮想通貨ドメインのインフルエンサー（仮想通貨インフルエンサー）と、仮想通貨ドメインに関連付けられていない GitHub 全体でのインフルエンサー（GitHub インフルエンサー）である。仮想通貨ドメインには両方のタイプのインフルエンサーが含まれている (図 3.3)。入次中心性、PageRank、HITS (hypertext induced topic selection) の 3 つのアルゴリズムを比較し、仮想通貨インフルエンサーを識別するために適切なアルゴリズムを選択する。これらのアルゴリズムは、WebGraph（Web ページのリンク関係に基づくグラフ）内で重要なサイトをランク付けするために考案されたもので、主に有向グラフのネットワーク分析に使用されている。これらのアルゴリズムについては、3.4 章で詳しく説明する。

RQ 2： インフルエンサーは他の貢献者よりも多く貢献活動しているか？

これは、ソーシャルメディアのインフルエンサーでは頻繁に研究されているリサーチクエスチョンである。本研究では仮想通貨ドメインで本件を検証し、従来の研究との差異を確認する。

GitHub に関する従来の研究では、貢献量が影響力に影響を与えないことが示されている [BSG+16, LRM14]。対照的に、Twitter のインフルエンサーになるには、頻繁にツイートするなどの相当な努力が必要であるとしている [CHB+10]。本研究では、GitHub をデータセットとして使用しているので、前者と同じ結果が予想さ

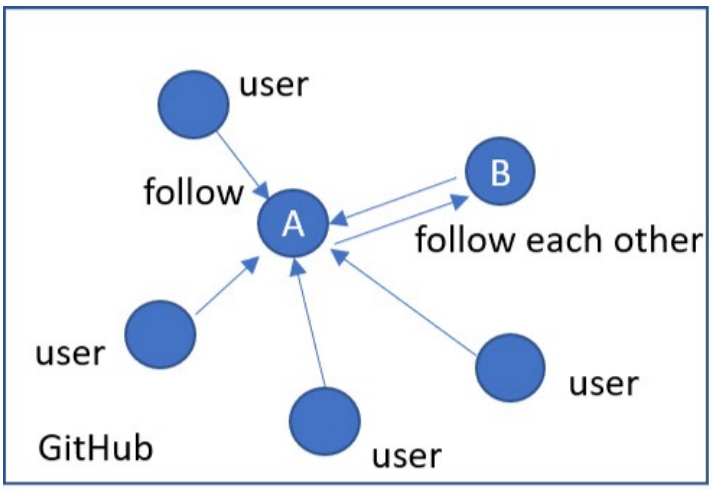

図 3.1　Follow-network

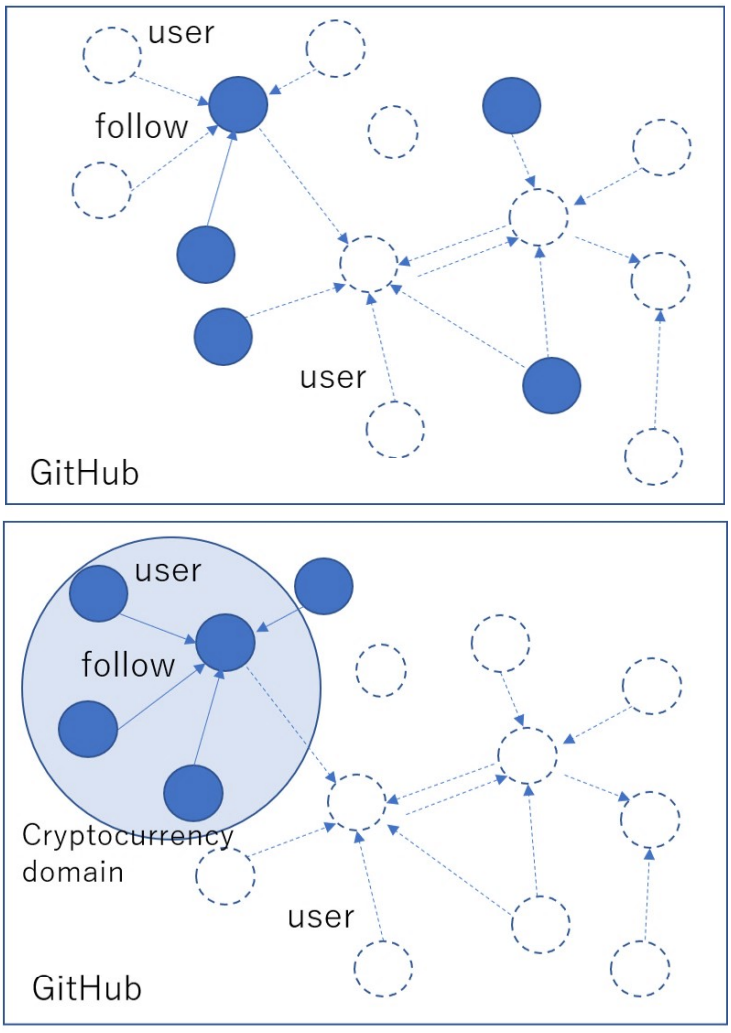

図 3.2　Random extraction (upper) and domain extraction (lower)

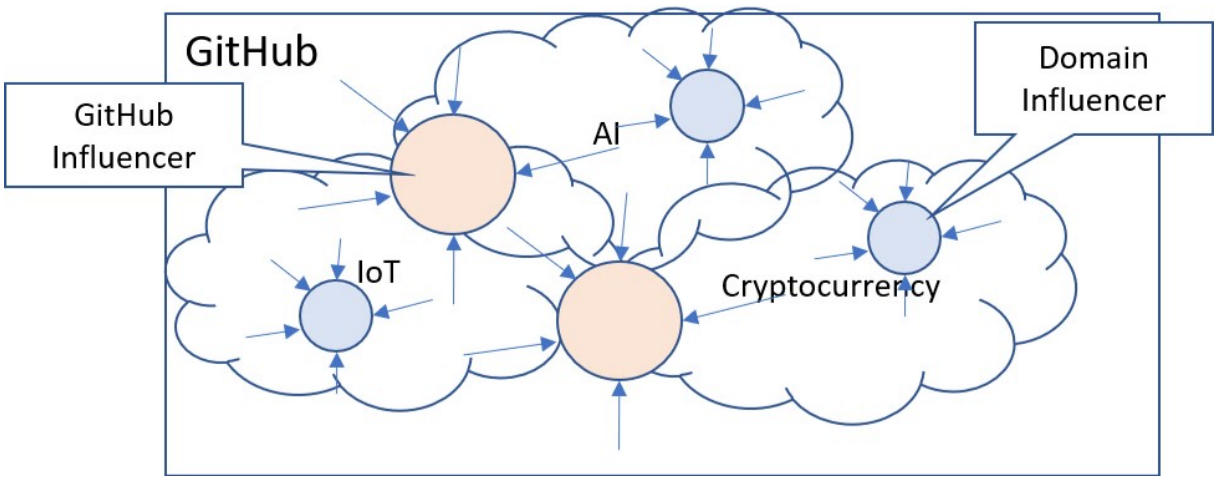

図 3.3　GitHub Influencer and Domain Influencer

れる。

しかしながら、仮想通貨プロジェクトは最先端の技術を使用しており、ユーザーが開発方法に興味を持ち、多大な貢献活動をしている開発者をフォローするなども考えられる。貢献量と影響力の関係を検証し、そのような結果が得られれば仮想通貨は GitHub の中で特殊なドメインであると言える。

RQ 3： インフルエンサーにより、貢献者をより多く獲得できるか？

Blincoe らは、GitHub において 500 人を超えるフォロワーを持つインフルエンサーを分析し、インフルエンサーが、フォロワーをプロジェクトに導くとした [BSG+16]。たとえば、インフルエンサーが新しいプロジェクトに貢献した場合、フォロワーの 13.7% がそのプロジェクトに貢献していた。また、インフルエンサーのフォロワー数と貢献したフォロワーの人数との関連性も調査した。ただし、ドメインインフルエンサーと GitHub インフルエンサーを区別していない。

本章は、次のように構成されている。3.2 章では、分析で使用されるデータセットとそのクレンジング処理について説明する。3.3 章では、仮想通貨のフォローネットワークの構造的特徴を、Twitter と GitHub のフォローネットワークに関する以前の研究結果と比較して説明する。3.4 章〜 3.6 章では、リサーチクエスチョンに対して回答する。最後に、3.7 章では、結論を述べ、将来の研究について提言する。

3.2 データセット

GitHub は、プロジェクトの情報を取得する API を提供している。この API を利用して、仮想通貨プロジェクトの貢献者数、コミット数、および活動期間を取得した。仮想通貨プロジェクトのリストは、仮想通貨市場ランキングチャートの Web サイトから取得した [*2]。 上記により、554 のプロジェクトと 2,444 の貢献者に関するデータを取得した。

同じ API を使用して、各貢献者のフォロワーデータを取得した。貢献者とフォロワーの総数（ユーザーの総数）は 70,217 で、フォローによるリンク数は 129,841 であった。多くのユーザーが貢献者とフォロワーの両方として活動していた。

次にデータのクレンジングをおこなった。

[*2] https://www.coingecko.com（accessed 2018-02-12）

1. ビットコインなどの主要な仮想通貨のソースコードをフォークまたは再利用して、いくつかのプロジェクトが派生的に作成されている。そのようなプロジェクトでは元のプロジェクトの貢献者とその活動履歴も継承される。つまり元のプロジェクトと活動履歴が重複し、一方のデータを削除する必要がある。派生プロジェクトを調査し、重複した貢献者とそのコミット履歴を削除した（図 3.4）。円のサイズは、貢献者の数を示している。クレンジング前は、ソースとデスティネーションのプロジェクトの貢献者数はほぼ同じである（左）。クレンジング後、デスティネーションプロジェクトの貢献者数が大幅に減少していることがわかる（右）。
2. GitHub の API は、すべての活動データを提供するわけではない [CLC16]。API を使用して取得できる貢献者の最大数は、プロジェクトごとに 100 である。 2017 年 3 月以降、貢献者の数は急激に増加し、一部のプロジェクトは 100 を超えている。このデータは 2018 年 2 月に取得されたもので、コミット数が少ないマイナーな貢献者は欠落している。主要な貢献者は網羅されており、本研究の結論に影響を与えたとは考え難い。
3. 一部の貢献者は複数のユーザー ID を使用している [VPR+15]。ユーザーを GitHub の属性（たとえば、ログイン名、実際の名前、メールアドレス、場所）を利用して同一人物か確認するツール [*3]を使用してデータを整理した。2,444 人の貢献者を調査し、そのうちの 10 人が重複していると判断し、重複した ID をマージした。その結果貢献者の総数は 2,434 となった。

図 3.5 は、貢献者間のフォロー関係を示している。ここで、ノードは貢献者であり、有向エッジはフォロワーからフォローされているユーザーへのフォロー関係を示している。色は、貢献者が主に属しているプロジェクトであり、緑：ビットコイン、橙：イーサリアムである。ノードサイズはフォロワー数を示している。ビットコインやイーサリアムなどの主要通貨が大きなクラスターを形成していることがわかる。エッジのないノードはフォロー機能を使用していない貢献者である。つまり、貢献者の数の約半分はフォロー機能を使用していない。

[*3] https://github.com/bvasiles/ght_aliases（accessed 2018-05-12）

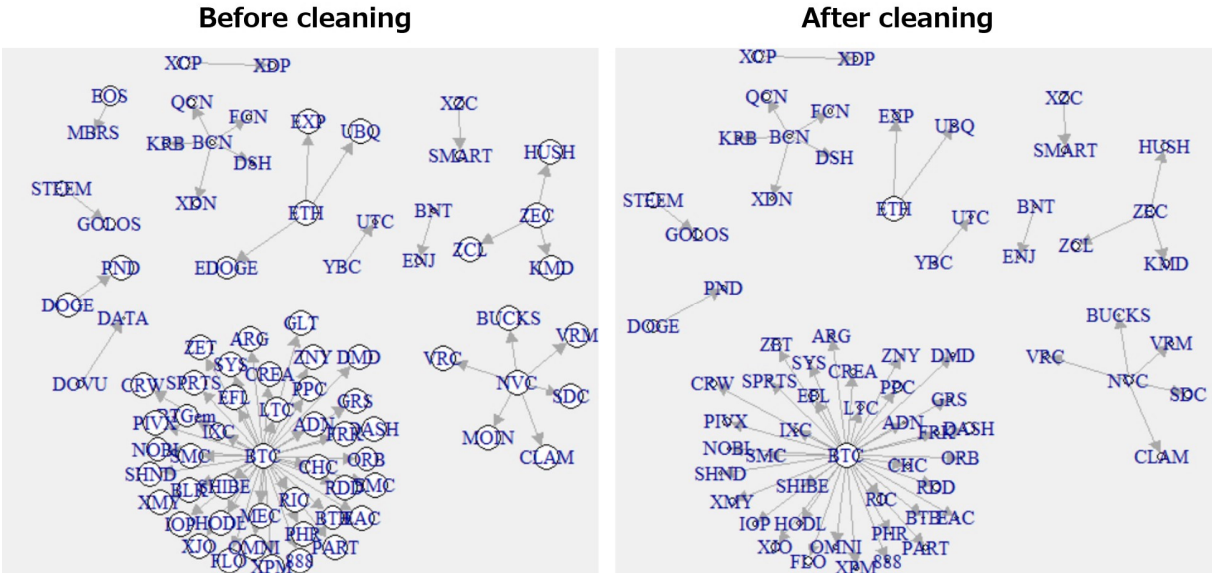

図 3.4　The cleaning process of forked projects (only projects with fork relationships are shown). (This data was obtained on 2018-02-12)

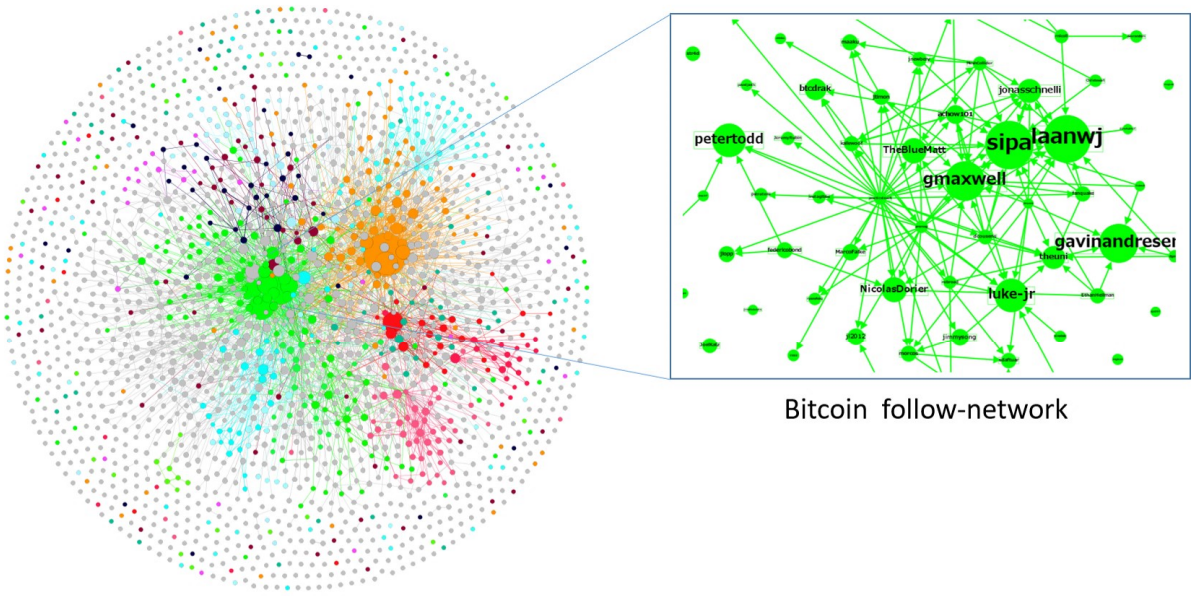

図 3.5　Cryptocurrency follow-network.(This data was obtained on 2018-02-12)

3.3 仮想通貨フォローネットワークの構造的特徴

リサーチクエスチョンに進む前に、この節では、仮想通貨フォローネットワークの構造的特徴を、先行研究で分析に使用された GitHub フォローネットワークおよび Twitter のフォローネットワークと比較して説明する。

平均次数とは、全フォロー数をユーザーの総数で割ったものである。仮想通貨フォローネットワークの平均次数は 1.85 で、先行研究の GitHub フォローネットワークでの値 0.74-1.01 [YYWW14] と 3.02 [LRM14] の中間に位置する。Yu らの研究では、GitHub からサンプリングされたユーザーによるサブセットを使用している。前述したように、その場合、始点ノードまたは終点ノードの一方しか抽出されない可能性が高く、ノード間のリンクが損なわれていると考えられる。つまり、一部のリンクがデータから除外されたため、平均次数が減少した。対照的に、Lima らはフォローイベント (リンク）からネットワークを構築した [LRM14]。つまり、データにはフォロー機能を使用するユーザーのみが含まれていた。各ユーザーの最低次数は 1 となり、平均次数を押し上げた。

GitHub での平均次数は Twitter の 35.2 [KLPM10] よりもはるかに小さい。 Twitter は情報サービスサイトであり、「フォロー」は中心的な機能であるが、GitHub はソフトウェアの共同開発サイトであり、「フォロー」は二次的な機能である。したがって、フォロー機能を使用する割合は低い。したがって、平均次数は減少する。また、後述するように、相互フォローの比率が低いことも影響した。

相互フォローしている貢献者の割合は 13.6% で、Twitter の 22.1% [KLPM10] よりも低かった。次の節で説明するように、上位インフルエンサーはごくわずかなユーザーをフォローしている、いわゆる「ロックスター」状態にある。たとえば、イーサリアムの作成者である vbuterin は 6000 人以上のフォロワーがいるが、誰もフォローしていない。

図 3.6 は、今回用いたデータセットの入次数と出次数の対数分布である。出次数の分布は直線に近くべき乗則分布に従っている。勾配の指標であるスケーリングインデックスは 2.4 である。入次数分布は、10 未満の範囲で下方向にバイアスしている。実際のネットワークでは、このようなバイアスが低い範囲内で発生することがよく知られている [New05]。入次数のスケーリングインデックスも 2.4 だった。 WebGraph や Citation Network などの実際のグラフのスケーリングインデックスは 2 から 3 [CSN09] の間であ

るため、本研究のデータはこの範囲に収まっている。

3.4 RQ 1：仮想通貨インフルエンサーをどのように特定できるか？

GitHub でユーザーをフォローすることは、ユーザーのプログラミングスキルと活動に興味があることを意味する [DSTH12]。フォロワーはフォローしているユーザーの影響を受けると推測される。多数フォローされているということは、かなりの影響力があることを示している（つまり、影響力のあるユーザーである）。この量的尺度は、入次中心性、つまりフォロワー数である。また、影響力のあるユーザーがフォロワーとして加わっている場合、影響力のあるユーザーを通じて間接的に多数のユーザーに影響を与えるため、影響力が高まることが想定される。 PageRank と HITS は、間接的な影響を考慮した中心性アルゴリズムである。これらのスコアは 0 ～ 1 の範囲に正規化されている。

PageRank アルゴリズムと HITS アルゴリズムの概要は次のとおりである。

- PageRank は Brin および Page [*4]によって考案されたアルゴリズムであり、リンク構造を分析して Web ページの重要性を定量化する。特定のページ A のページランク $PR\left(A\right)$ は、次の方程式で定義される。

$$PR(A) = (1-d) + d\sum_{i=1}^{n} \frac{PR\left(T_i\right)}{C\left(T_i\right)} \tag{3.1}$$

  ここで、n はページ A にリンクしているページ数である。 $PR\left(T_i\right)$ は、ページ T_i のページランクである。$C\left(T_i\right)$ は、ページ T_i から他のページへのリンク数である。d は、Web サーファーがサーフィンを続ける確率である（減衰係数とも呼ばれる）。通常は 0.85 に設定される。

- HITS も RageRank と同様に Web ページを評価するアルゴリズムである [Kle99]。HITS アルゴリズムには、すべてのページに対する 2 つの中心性スコア（Authority スコアと Hub スコア）が存在する。

  Authority スコア $(auth(A))$ は、ページ A にリンクしているページの Hub スコアの合計である。

  Hub スコア $(hub(A))$ は、ページ A からリンクしているページの Authority スコア

[*4] http://infolab.stanford.edu/ backrub/google.html（2019-01-20 にアクセス）

の合計である。

各スコアは、反復計算によって更新される。

最初に、$\forall p$, $auth(p)$=1;$hub(p)$=1 に初期化する。

Hub スコアを合計して $auth(p)$ を更新する：

$$auth(p) = \sum_{i=1}^{n} hub(i) \tag{3.2}$$

次に $\forall p$、Authority スコアを合計して $hub(p)$ を更新する。

$$hub(p) = \sum_{i=1}^{n} auth(i) \tag{3.3}$$

反復計算ごとに両方の値の正規化が必要である。

前述したように、インフルエンサーには 2 つのタイプが存在する、仮想通貨インフルエンサーと GitHub インフルエンサーである。仮想通貨インフルエンサーを抽出するのに最適な中心性アルゴリズムを確認した。

表 3.1 は、各中心性スコアの影響力の高い上位ユーザーを示している。 GitHub の各ユーザーのプロファイル、ブログを調査し、そのコンテンツに仮想通貨に関連した文言（たとえば仮想通貨名）が含まれているか確認した。また、同じランクのインフルエンサーとフォロワー数が同等か確認した。表 3.1 の太字のユーザー ID は、仮想通貨インフルエンサーとは言い難いユーザーである。 Authority スコアの上位ユーザーはすべて仮想通貨インフルエンサーであるが、入次中心性と PageRank の上位ユーザーはさまざまなタイプのユーザーが混在している。入次中心性には、仮想通貨との関連性は低いがフォロワーが非常に多い GitHub インフルエンサーが含まれる。上記に加えて、PageRank にはフォロワー数の少ないユーザーも含まれていた。

Authority スコアが高いユーザーの属性を表 3.2 に示す。各ユーザーの出次数は非常に小さい。つまり、影響力のあるユーザーが他のユーザーをフォローすることは希少である。 PageRank スコアはフォロワーのスコアを出次数で割った合計であるため、非常に大きなスコアと小さな出次数を持つフォロワーは、フォローしているユーザーの PageRank スコアを大きく増大させて、全体のランクを大きく歪めてしまう。たとえば、AndrewScheidecker には 181 人のフォロワーがいるが、これは影響力のあるユーザーと比べかなり少ない。ただし、2506 人のフォロワーを持つ bytemaster が

表 3.1　Top 10 influential users ranked based on centrality scores

	In-degree	**PageRank**	**HITS/Authority**
1	vbuterin	vbuterin	vbuterin
2	***soulmachine***	DavidVorick	laanwj
3	bytemaster	sipa	sipa
4	laanwj	***soulmachine***	gmaxwell
5	***graydon***	laanwj	gavinandresen
6	gavinandresen	bytemaster	luke-jr
7	***jonathanong***	lukechampine	petertodd
8	sipa	***AndrewScheidecker***	jonasschnelli
9	***jedisct1***	***chrisdone***	bytemaster
10	gmaxwell	gmaxwell	gavofyork

AndrewScheidecker だけをフォローしているため、bytemaster のすべてのスコアが彼に割り当てられ、PageRank で 8 番目に影響力のあるユーザーとなっている。

図 3.7 は、PageRank には、フォロワー数は少ないものの高いスコア（青い円内）を持つ多くの同様なユーザーが存在することを示している。 Authority では、そのようなユーザーは見られない。仮想通貨フォローネットワークには、複数の仮想通貨インフルエンサー（N) をフォローする、仮想通貨プロジェクトに関心をもつ多くのユーザー（M) が含まれる。いわゆる N 対 M の関係が構成されている。この場合、それぞれのユーザーの Authority スコアと Hub スコアが反復計算により大幅に増加する。 GitHub インフルエンサーには多くのフォロワーがいるが、関係は 1 対 M である。この場合、どちらのスコアも大幅に増加しない。このため、Authority スコアの上位ユーザーはすべて仮想通貨のインフルエンサーである。

結論として、HITS は、仮想通貨フォローネットワークからそのドメインのインフルエンサーを抽出するのに適したアルゴリズムである。 HITS は、さまざまなタイプのインフルエンサーが混在するフォローネットワークでドメインインフルエンサーを抽出するのに適したアルゴリズムであると考えられる。

表 3.2　Attributes of top 10 influential users ranked by Authority scores

	user ID	In-degree	Out-degree	Score	**Currency**
1	vbuterin	6472	0	1.000	ETH(Etherium)
2	laanwj	2033	1	0.309	BTC(Bitcoin)
3	sipa	1493	3	0.269	BTC(Bitcoin)
4	gmaxwell	1133	0	0.220	BTC(Bitcoin)
5	gavinandresen	1590	0	0.213	BTC(Bitcoin)
6	luke-jr	980	0	0.198	BTC(Bitcoin)
7	petertodd	1089	0	0.189	BTC(Bitcoin)
8	jonasschnelli	671	3	0.156	BTC(Bitcoin)
9	bytemaster	2506	1	0.153	BTS(Bitshare)
10	gavofyork	1026	0	0.143	ETH(Etherium)

3.5　RQ 2：インフルエンサーは他の貢献者よりも多く貢献活動しているか？

本節では、ユーザーの影響力とユーザーの貢献量との関係を調べた。Authority スコアを影響力の指標として使用した。コミット数、活動期間、および参加プロジェクト数は、貢献量の変数である。参加プロジェクト数を除く各変数は、正規分布に近づくように自然対数スケールに変換した。データセットはトレーニング用とテスト用にランダムに分割した。著名な機械学習手法 [K$^+$08] を用いトレーニングセットで学習した。次にテストセットを使用して検証をおこない、その残差を取得した。この手順を 5 回繰り返し、平均して結果を得た。

結果を表 3.3 に示す。 二乗平均平方根誤差（RMSE）最小二乗法 (OLS) の最小だったが、各手法の値は極めて近かった。各決定係数 (R-Squared) は 0.036 から 0.083 までと極めて小さかったため、各貢献量の変数は影響力の予測には有効ではない。図 3.8 は、SVM の予測結果に対して Authority スコアをログスケールでプロットしている。貢献量による影響力への予測力が高い場合、各データポイントは破線に沿って分布する必要がある。こ

表 3.3　RQ2: Results of machine learning prediction

	Performance across resamples		Variable Importance		
	RMSE	R-Squared	Log(Commit)	Log(Day)	Project
OLS	2.39	0.083	87	100	0
Random Forest	2.47	0.058	63	100	0
SVM	2.41	0.085	-	-	-
XGBoost	2.66	0.036	77	100	0

れらのデータポイントはプロット内でランダムに散在しており、そのような傾向は観察されなかった。

図 3.9 は、仮想通貨フォローネットワーク全体を示している。ここで、ノードは貢献者、有向エッジは、フォロー関係を示している。色は、主に所属するプロジェクトであり、緑 : ビットコイン。橙 : イーサリアムである。ノードのサイズは、各ユーザーの Authority スコア（上）とコミット数（下）を表す。 Authority スコアの図では、上位スコアのユーザーは大規模プロジェクトの中心に集まっている。一方、コミット数の多いユーザーは多数のプロジェクトに散在している。図 3.9 は、多数コミットしているユーザー、すなわち貢献活動を積極的におこなっているユーザーと影響力を持つユーザーが同一でないことを視覚的に示している。

仮想通貨のドメインにおける検証結果は、GitHub 全体での先行研究の結果「積極的に貢献活動しているユーザーが必ずしも多数のフォロワーを持っているわけではない」[LRM14] と一致している。

3.6　RQ 3：インフルエンサーにより、貢献者をより多く獲得できるか？

影響力のあるユーザーがプロジェクトに参加している場合、プロジェクトへの貢献者を増やすことはできるか？この疑問に答えるために、プロジェクト内で最も影響力の高いユーザーの Authority スコアがプロジェクトへの貢献者数と関連しているか分析した。重回帰分析を適用して、Authority スコア（auth)、およびリポジトリのサイズ（size)、ス

ター数 (stars)、プロジェクト存続日数（life)、フォーク数 (forks) などのプロジェクトの属性に基づいて、貢献者数を予測した。Authority スコアなし（Model1）と Authority スコアあり（Model2）の 2 つのモデルを作成して効果を比較した。

分析の前に正規分布への適合性をチェックし、auth、size、stars、forks、および貢献者数を対数変換した。多重共線性は、分散拡大係数（VIF）を使用して評価した。 10 を超える VIF は深刻な多重共線性が存在すると見なされ、4.0 を超える値は多重共線性の存在が疑われる [HJARTRBW94]。stars と forks の VIF はそれぞれ 5.4 と 5.7 であった。したがって、1 つの変数（forks）を削除したモデル (Model0) を作成し比較対象とした。

表 3.4 に重回帰分析の結果を示す。 Model0 と Model1 の RMSE 値は同じであるが、Model1 の自由度調整済み決定係数 (Adj.R^2) と赤池情報量基準（AIC）は、Model0 よりも優れていた。したがって、forks を含むモデル、つまり Model1 を採用した。

Model1 と Model2 の両方においてすべての説明変数は有意であった。Adj.R^2、RMSE および AIC は、auth を追加することで改善された。念のため両方のモデルの差の存在を検証するため分散分析を実施した。結果は、モデルの差が存在しない確率が 0.1 %有意であり、モデルの差の存在が確認された。

これは、「影響力のあるユーザーがプロジェクトに貢献者を誘導する」という GitHub に関する先行研究の結果と一致している [TDH13, BSG+16]。

3.7 本章のまとめ

本研究では、GitHub 上の仮想通貨プロジェクトのフォローネットワークを定量的に分析し、以下を発見した：

- プロジェクトインフルエンサーの影響力は、貢献者数と関連性がみられる。これは、影響力のあるユーザーがプロジェクトへの貢献者を集めることに貢献していることを示唆している。
- 仮想通貨プロジェクトには、一般的な GitHub インフルエンサーと仮想通貨、つまりドメイン固有のインフルエンサーの両方が存在する。上位インフルエンサーは「ロックスター」のような存在であり、多くのフォロワーを集めるが、自分自身はほとんどフォローしない。
- 3 つの中心性スコア（入次数、PageRank、HITS/Authority）を比較することによ

表 3.4　Summary of hierarchical multiple regression analysis for variables predicting contribution (N=312)

	Model0	Model1	Model2
(Intercept)	−1.02 (0.23)***	−1.06 (0.23)***	0.24 (0.30)
size	0.10 (0.03)***	0.09 (0.03)***	0.07 (0.02)**
stars	0.35 (0.02)***	0.25 (0.05)***	0.19 (0.05)***
life	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)***
forks		0.15 (0.07)*	0.13 (0.06)*
auth			0.15 (0.02)***
R^2	0.54	0.54	0.60
Adj. R^2	0.53	0.54	0.59
Num. obs.	312	312	312
RMSE	0.77	0.77	0.72
AIC	728.4	725.6	687.2

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

り、Authority スコアが、仮想通貨インフルエンサーの影響力の指標として最も適していることがわかった。 HITS アルゴリズムは、仮想通貨フォローネットワークなどのようなネットワークからドメイン固有のインフルエンサーを抽出するのに有効と考えられる。

- ユーザーの貢献量は、影響力との関連性は見られなかった。インフルエンサーになるために多大な活動が必要な Twitter と比べ、GitHub では貢献活動量は重要ではないと結論づけられる。

入次中心性と PageRank アルゴリズムは、影響力の指標として頻繁に使用されている。ただし、これらのアルゴリズムは、フォロワーの間接的な影響（入次中心性）を全く考慮しないか、フォロワーの間接的な影響を過度に考慮してしまう（PageRank アルゴリズム)。「ロックスター」を含むネットワークでは、HITS アルゴリズムが、これらのアルゴリズムよりも優れていることがわかった。今回は仮想通貨フォローネットワークを分析したが、HITS アルゴリズムは他のドメインでも効果的であると考えている。「ロックスター」の存在は他の研究 [DSTH12, LFCH13, TDH14] で頻繁に報告されており、GitHub の一般的な構造である可能性が高い。

仮想通貨は最先端のテクノロジーで構成されており、プロジェクトの成果物は開発者にとってより関心が高いと考えられる。一般的なプロジェクトでは多大な貢献活動をした開発者だからといって、フォローされるわけではなかったが、仮想通貨プロジェクトでは異なった結果になるのではないかと期待していた。しかしながら、貢献量と影響力の間には関連性がみられなかった。先行研究で示されたように、一般的なプロジェクトと同様に他のユーザーをフォローする基準は、プロジェクトに関する他の情報、またはブログやソーシャルメディアなどの外部活動に依存すると推測される。

本章と次章で使用した仮想通貨のデータセットは以下の点でユニークであり、今後さまざまな分析に適用可能と考えられる。

1. GitHub 上の仮想通貨開発プロジェクトを抽出している。つまりプロジェクトは仮想通貨という共通ドメインに所属しており互いに関連している。過去の研究では関連性のないプロジェクトの集合を分析していた。
2. 仮想通貨には、市場価格や取引量（ボリューム）などの金融情報が存在する。またソーシャルメディアでの情報、検索エンジンでの検索数、決済件数、新聞記事など、プロジェクト以外の多様な情報が存在し、それらと組み合わせて分析が可能である。

OSS の貢献者に関しては、多くのユニークな研究が存在する。これらの研究手法を本データセットに適用することが可能である。たとえば、山下ら [YMKU14, YKM+16] は、マグネットとスティッキーのメトリックを使用して、「貢献者を維持するプロジェクト」、「貢献者を獲得するプロジェクト」に分類した。本データセットをこれらのメトリックと組み合わせることにより、貢献者の行動をより詳細に分析できると確信している。

次の章では、本章と同様仮想通貨のデータセットを用い、時価総額と貢献者数の関連性について分析をおこなう。

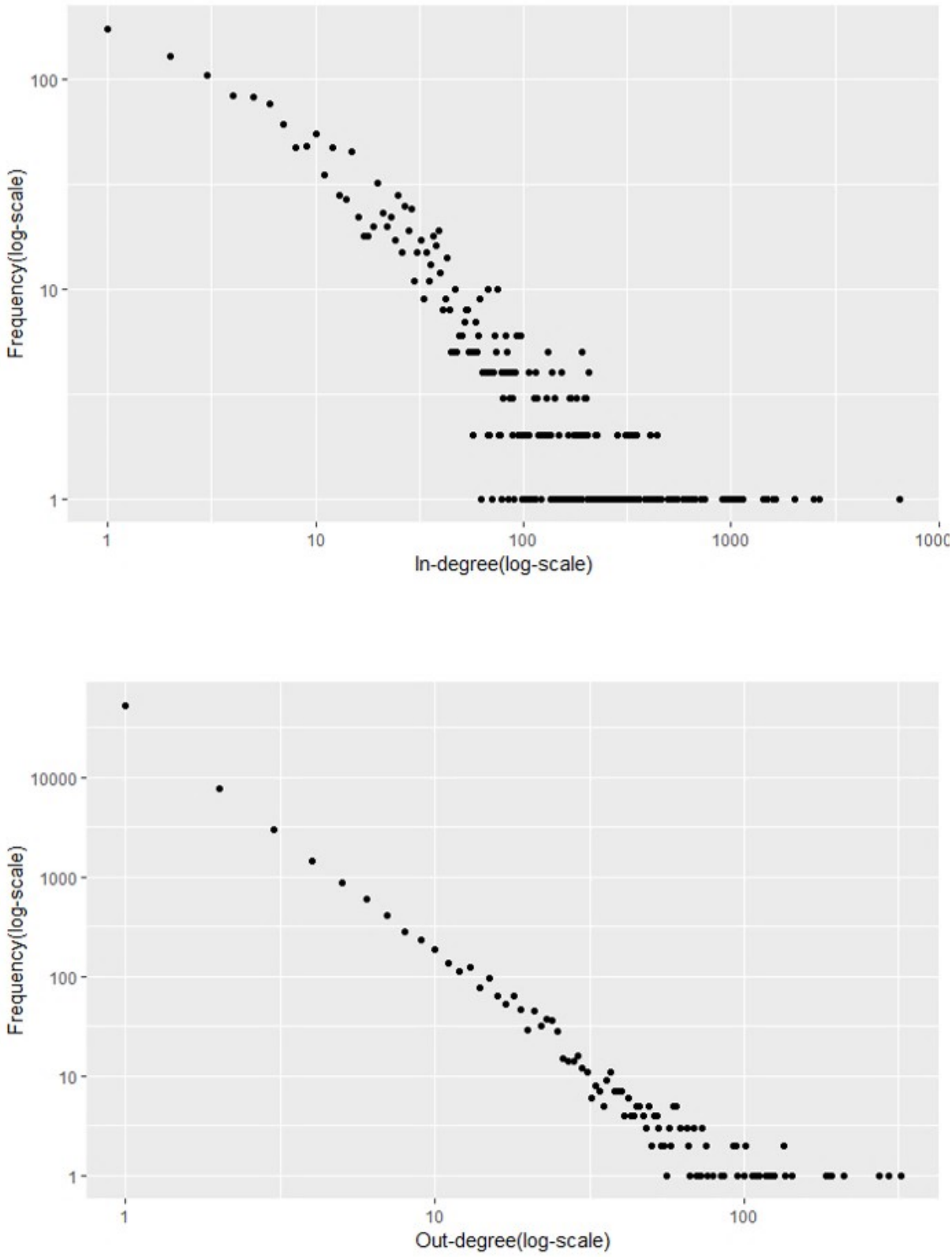

図 3.6　Follow-network degree distribution; in-degree (upper), out-degree (lower).

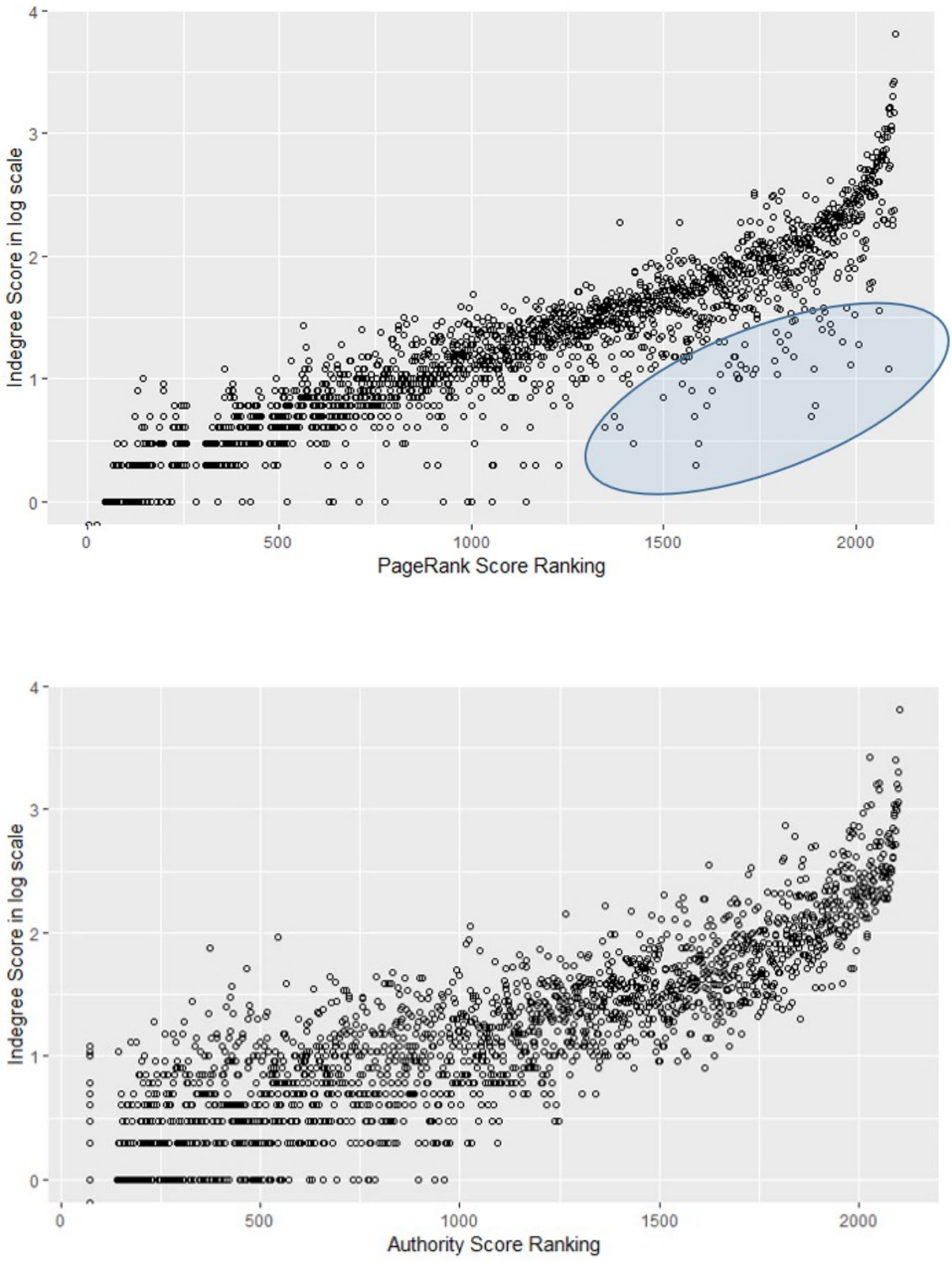

図 3.7　Score ranking of PageRank (upper) and Authority (lower).

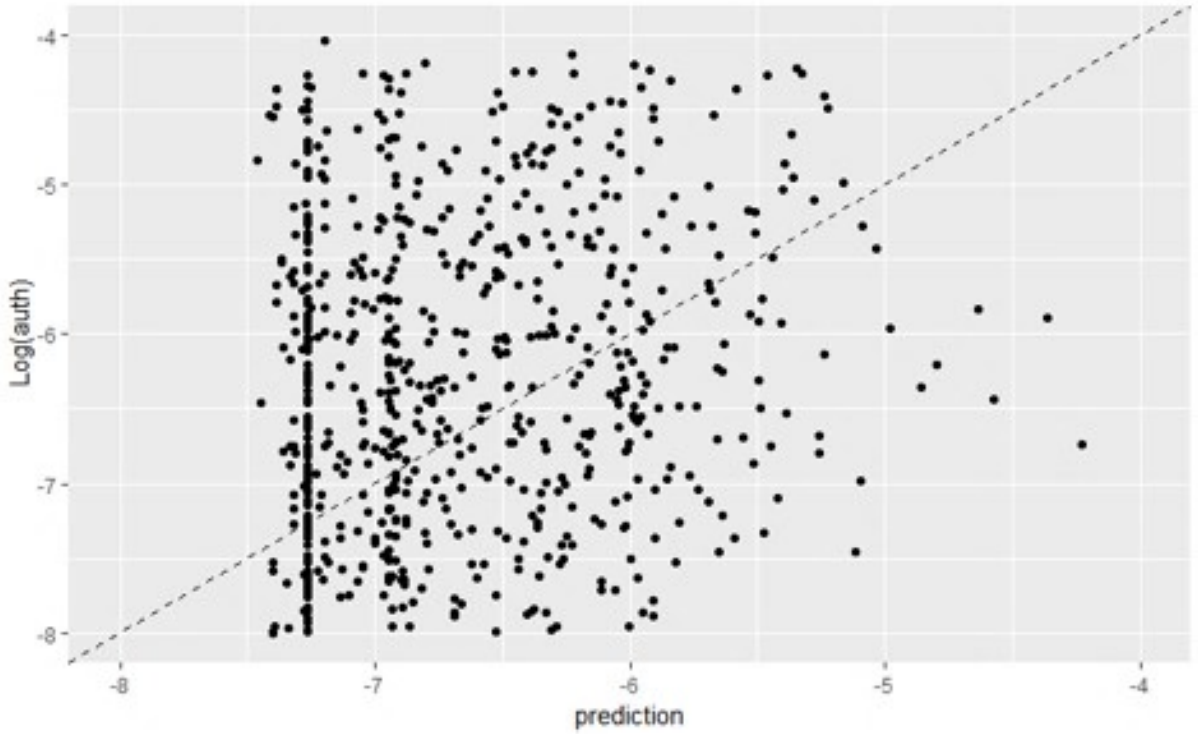

図 3.8　Prediction using SVM

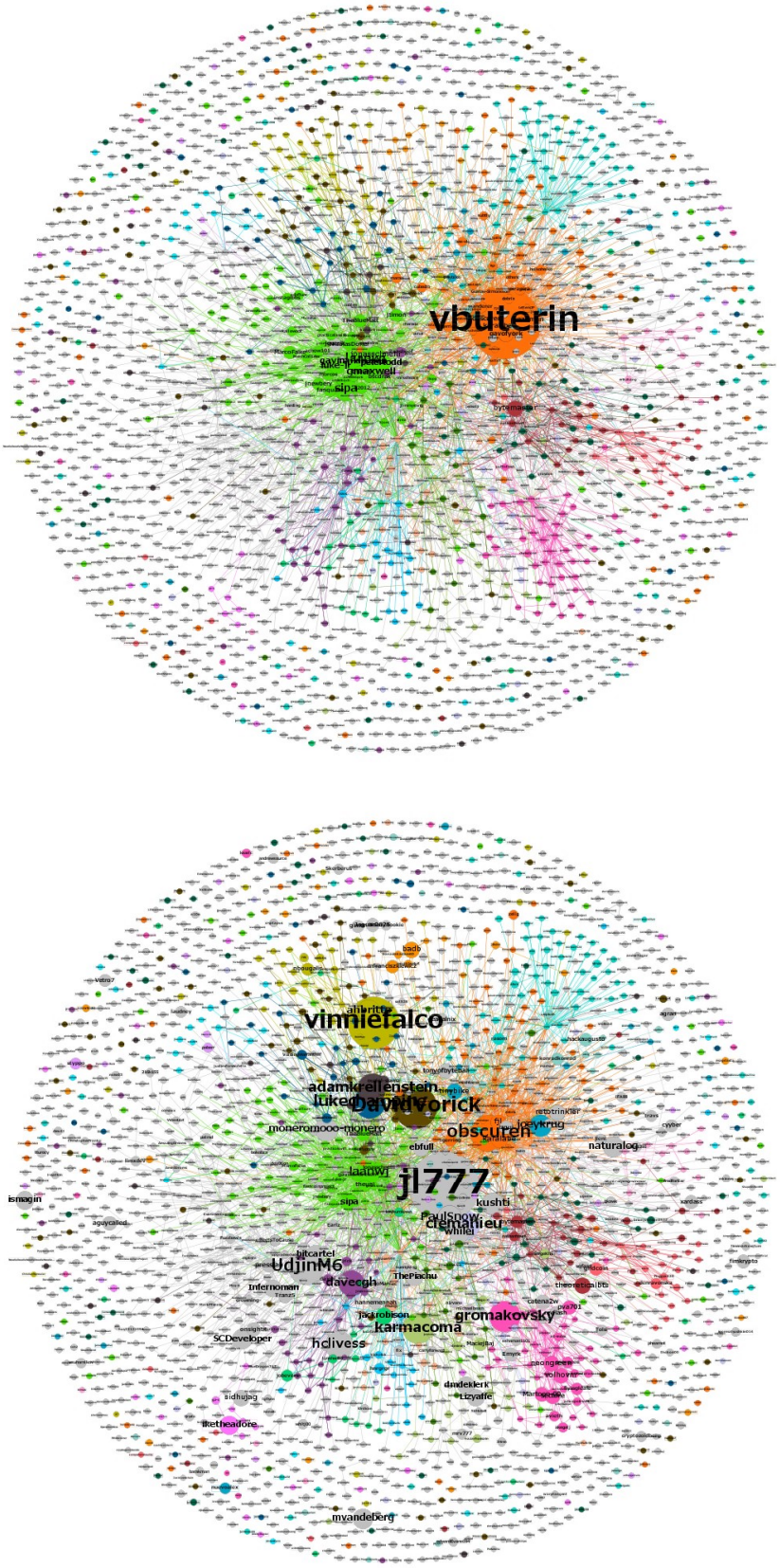

図 3.9　Cryptocurrency follow-network. The node size indicates the Authority score (upper) and the number of commits (lower). (This data was obtained on 2018-02-12)

第 4 章

時系列分析による仮想通貨の時価総額と貢献量との関連性の研究

本章では、貢献者獲得の要因としてプロジェクトの将来性について分析をおこない、有効性を確認する。前章と同じく、分析データとしては GitHub 上の仮想通貨プロジェクトを使用する。仮想通貨は市場データが公開されており、時価総額の推移が取得できる。時価総額をプロジェクトの将来性を示す代理変数として、貢献者数との関連性を時系列分析手法で検証する。

4.1 序論

将来性のあるプロジェクトに貢献することは、ユーザーの評判やキャリアにとって有利に働くであろう、一方、プロジェクトに将来性がなければ、彼らの貢献はすべて無駄になってしまうかもしれない。Dabbish らはインタビューによる調査の結果、ユーザーがどのプロジェクトが長期的に成功するか、公開されている情報を基に判断していることを発見した [DSTH12]。プロジェクトの将来性は、貢献者獲得につながる重要な要因と推測される。ただし、プロジェクトの将来性と、貢献者数の因果関係はまだ研究されていない。

本章では、時価総額を将来性の代理変数として採用して、貢献者数との関連性をグレンジャー因果性検定と回帰による時系列分析手法を使用して分析した。時価総額とは、上場企業の株価に発行済株式数を掛けたものであり、企業価値を評価する際の指標である。時価総額は入手可能なすべての情報を反映する [MF70]。 つまり、時価総額には、現在の業績だけではなく将来の成長に対する期待も含まれている。そのため、企業の真の価値の一

般的な指標としてよく使用されている [ZS10, Abd05] [*1]。仮想通貨の時価総額は、通貨の市場価格に流通発行量 (the total circulating supply) を掛けたものに等しい [*2]。仮想通貨の時価総額が増加すると、仮想通貨は将来的に有望であると推測され、貢献先に決定される可能性が高くなると想定される。

上記を踏まえ、次のリサーチクエスチョンを提案した。

RQ： ユーザーが、貢献先を決定する際、将来性は重要な要因となるのか？言い換えれば、将来性の指標である時価総額は、貢献者の OSS 開発プロジェクトへの参加に影響を与えるか？

GitHub から抽出した仮想通貨プロジェクトに対し、グレンジャー因果性検定 [Gra69] と回帰分析を適用して時価総額と貢献者数の因果関係の存在を調査した。

この章の構成は次のとおり。

4.2 章では、仮想通貨の時価総額の取得方法を説明して、そのクレンジングについて説明する。 GitHub 上の仮想通貨プロジェクトの時間的特徴については、4.3 章で説明する。4.4 章では、上記のリサーチクエスチョンに対する結果を示す。4.5 章で、検証の結果をまとめる。

4.2 データセット

冒頭で述べたように、仮想通貨の時価総額と、仮想通貨プロジェクトの貢献者数の時系列データを用い分析をおこなった。データは下記手順で取得した。

1. 市場ランキングチャートの Web サイトから仮想通貨のリストを取得した [*3]。このサイトは、2013 年 4 月以降の時価総額（米ドル）の時系列データおよび開発環境の URL を提供する。時価総額上位 30 の仮想通貨（表 4.1）を含む 584 の仮想通貨が GitHub を開発環境として使用していた。
2. GitHub を開発環境としている 584 の URL にアクセスを試みた。一部の URL は存在しなかったが、合計 554 のプロジェクトにアクセス可能であった。GitHub API

[*1] https://ja.wikipedia.org/wiki/時価総額（2020-04-15 にアクセス)
[*2] https://blockgeeks.com/guides/cryptocurrency-market-cap/（2019-10-04 にアクセス）
[*3] https://www.coingecko.com（accessed 2018-02-12）

表 4.1 Cryptocurrencies with the top 30 market capitalization
(source: https://www.coingecko.com (accessed 2018-02-12))

name	URL
Bitcoin (BTC)	https://github.com/bitcoin/bitcoin
Ethereum (ETH)	https://github.com/ethereum/go-ethereum
Ripple (XRP)	https://github.com/ripple/rippled
Bitcoin Cash (BCH)	https://github.com/Bitcoin-ABC/bitcoin-abc
Cardano (ADA)	https://github.com/input-output-hk/cardano-sl
Litecoin (LTC)	https://github.com/litecoin-project/litecoin
Stellar Lumens (XLM)	https://github.com/stellar/stellard
NEO (NEO)	https://github.com/neo-project/neo
EOS (EOS)	https://github.com/EOSIO/eos
NEM (XEM)	https://github.com/NewEconomyMovement/NemCommunityClient
Dash (DASH)	https://github.com/dashpay/dash
Monero (XMR)	https://github.com/monero-project/monero
Lisk (LSK)	https://github.com/LiskHQ/lisk
Ethereum Classic (ETC)	https://github.com/ethereumproject/go-ethereum
Qtum (QTUM)	https://github.com/qtumproject/qtum
ICON (ICX)	https://github.com/theloopkr/loopchain
Zcash (ZEC)	https://github.com/zcash/zcash
Steem (STEEM)	https://github.com/steemit/steem
Bytecoin (BCN)	https://github.com/amjuarez/bytecoin
Verge (XVG)	https://github.com/vergecurrency/verge
Status (SNT)	https://github.com/status-im/status-react
Siacoin (SC)	https://github.com/NebulousLabs/Sia
Stratis (STRAT)	https://github.com/stratisproject/Breeze
BitShares (BTS)	https://github.com/BitShares/bitshares-2
Aeternity (AE)	https://github.com/aeternity/epoch
Dogecoin (DOGE)	https://github.com/dogecoin/dogecoin
Veritaseum (VERI)	https://github.com/veritaseum/Veritaseum
Waves (WAVES)	https://github.com/wavesplatform/Waves
Augur (REP)	https://github.com/AugurProject/augur-core
0x (ZRX)	https://github.com/0xProject/contracts

を利用して、各プロジェクトの貢献者のコミット履歴を取得した。

下記手順でデータをクレンジングした。

1. 前章と同様、派生プロジェクトを調査し、重複した貢献者とそのコミット履歴を削除した。
2. 貢献者の存在しないプロジェクトは除外した。これらは単にソースコードレポジトリとして使用されていると想定される。その結果 457 プロジェクトが残った。

4.3 仮想通貨プロジェクトの時系列的な特徴

本章では、仮想通貨プロジェクトの記述統計、構造、および時間的変化について説明する。

前述のように、プロジェクトの数は 457 で、それぞれに 1 人以上の貢献者がいる。 うち 202 のプロジェクトでは、半年以上活動履歴がなかった。それらを「消滅」とみなすと、2018 年 2 月のプロジェクトの平均寿命は 287 日であった。活動履歴によると、一度以上コミットした貢献者数は 2434 人だった。一度だけしかコミットしなかった貢献者は全体の約 20%（504）である。この割合は、以前の研究で得られた 48.98% [PSG16] よりもはるかに低い。つまり仮想通貨プロジェクトには積極的に貢献しているユーザーが多いと推測される。

貢献者は複数のプロジェクトに参加している。貢献者の約 10%（254 人）が複数のプロジェクトに参加した。この値は先行研究の 4% より若干高いが [LAL+20]、対象プロジェクト数（先行研究は 297 の大規模プロジェクト）の差によるものと想定される。図 4.1 は、4 年間の仮想通貨のネットワーク構造の推移である。貢献者が同時期に参加したプロジェクト（円）は、辺でつながっている。円のサイズは、貢献者の数に比例する。指定された期間中に 6 人を超える貢献者がいるプロジェクトを表示している。赤い文字は「消滅」プロジェクトを示し、青い文字は時価総額が上位 10 位以内のプロジェクトを示す。ビットコインとその派生プロジェクトが OSS プロジェクトの大部分を占めていることがわかる。他のプロジェクトは毎年成長しており、複数のプロジェクトに参加する貢献者が増加してきていることがわかる。

図 4.2 は、毎年のプロジェクトの貢献者数（左）とプロジェクトの時価総額（右）の比

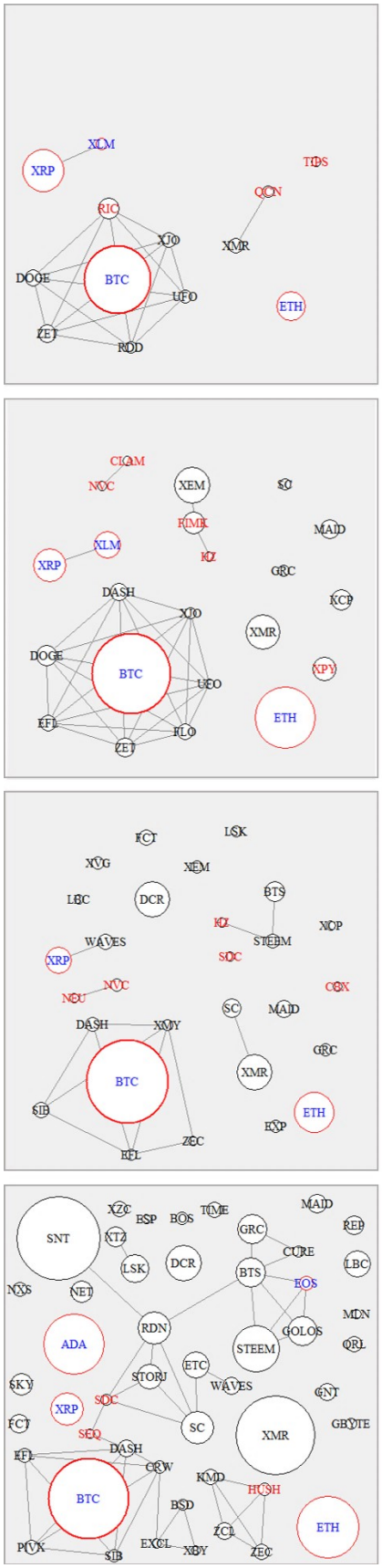

図 4.1　Network structures of cryptocurrency projects for each year.

率を示している。貢献者数に関しては、上位 10 プロジェクトの割合が徐々に低下している。ただし、長年にわたって極端な変化は見られない。ビットコインは、すべての期間の時価総額で圧倒的な割合を占めている。 2 位以降の順位は毎年入れ替わっている。ただし、年単位で大きな変化は見られない。仮想通貨の時価総額は単調に増加していない。プロジェクトを支える貢献者の数についても同様である。図 4.3 は、すべてのプロジェクトを合算した時価総額の時系列（MC として表示）および貢献者の週ごとの合計（WUC として表示）を示している。ソーシャルイベントは赤い点線で示した。全体の傾向は、いくつかのソーシャルイベントをきっかけに変化した。 2013 年 12 月、仮想通貨の時価総額の大きなシェアを占めるビットコインの価格は、中国政府の金融機関に対する規制導入により急落した（図 4.3 {1}）。それ以降時価総額は長期的に低迷した。2015 年 10 月、欧州司法裁判所がビットコインの売買について付加価値税（消費税）の対象ではないとの判決を下した（図 4.3 {2}）。それを境に時価総額は上昇し始めた。2017 年 3 月に時価総額は急激に上昇を始めた（図 4.3 {3}）。

4.4　時価総額とプロジェクト貢献者の分析

4.4.1　時価総額と貢献者数の関係

時系列を分析する前に、毎年の時価総額と個々の仮想通貨の貢献者との関係を概観する。図 4.4 は、年初の時価総額と年初から半年後の間に活動記録を持つ貢献者（Active Contributor）数との関係を示している。両方のデータは自然対数スケールに変換して、正規分布に近づけた。円のサイズは、期間全体の Active Contributor の合計を示す。プロジェクトの時価総額と Active Contributor の両方が、毎年徐々に増加していることが見て取れる、そして、より大きな時価総額を持つ仮想通貨プロジェクトがより多くの Active Contributor を持つ傾向が顕著に示されている。

4.4.2　グレンジャー因果性検定

時価総額は貢献者の行動に影響するのか、どの程度影響があり、いつ影響が発生するのか。これらの疑問に答えるために、全プロジェクトの時価総額の合計（MC として表される）および各週のユニークな貢献者数の合計（WUC として表される）に対し、時系列分析の手法であるグレンジャー因果性検定および回帰分析を実施した。ここで、MC と

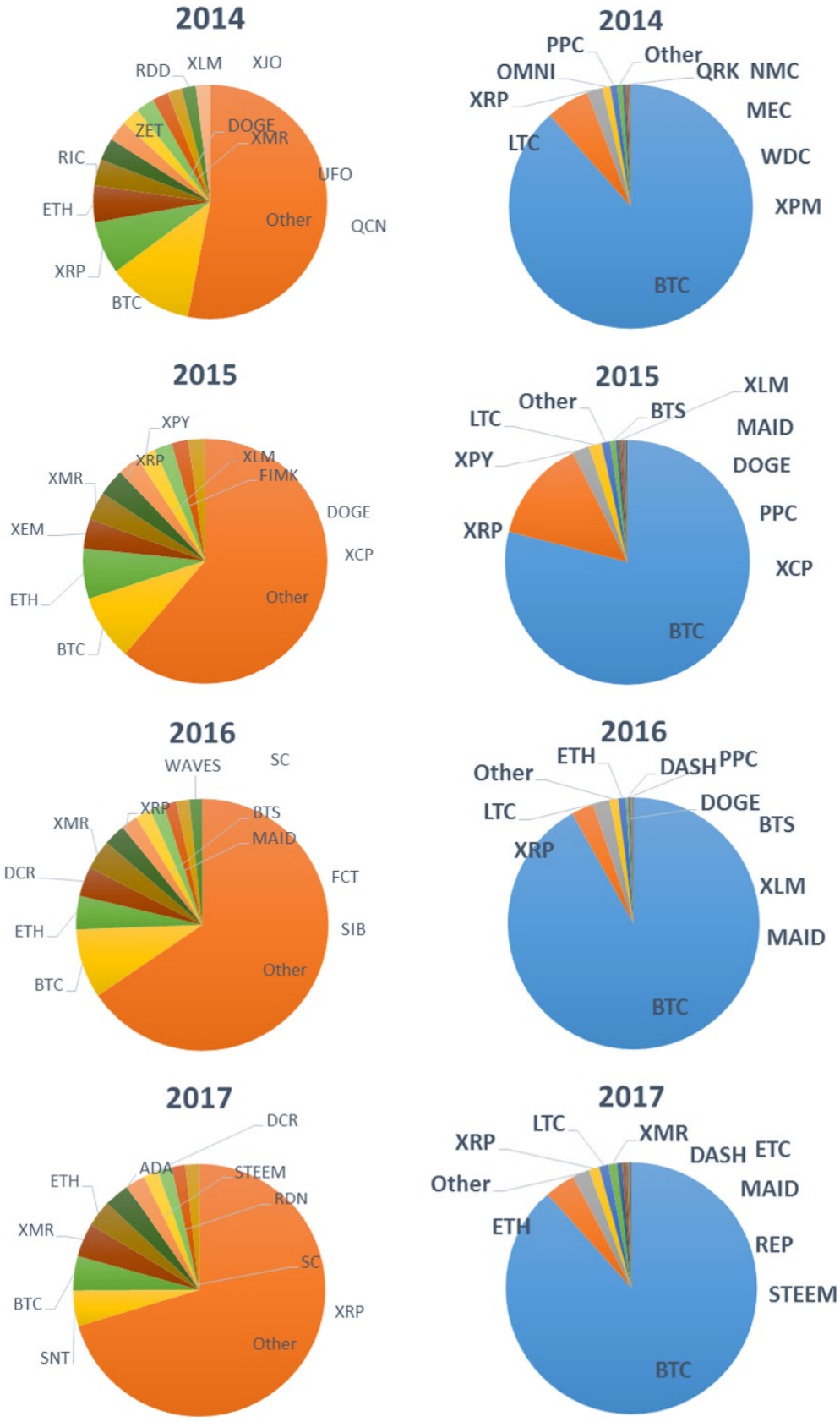

図 4.2　Percentage of the contributors (left) and the market capitalization (right) of the cryptocurrency for each year.

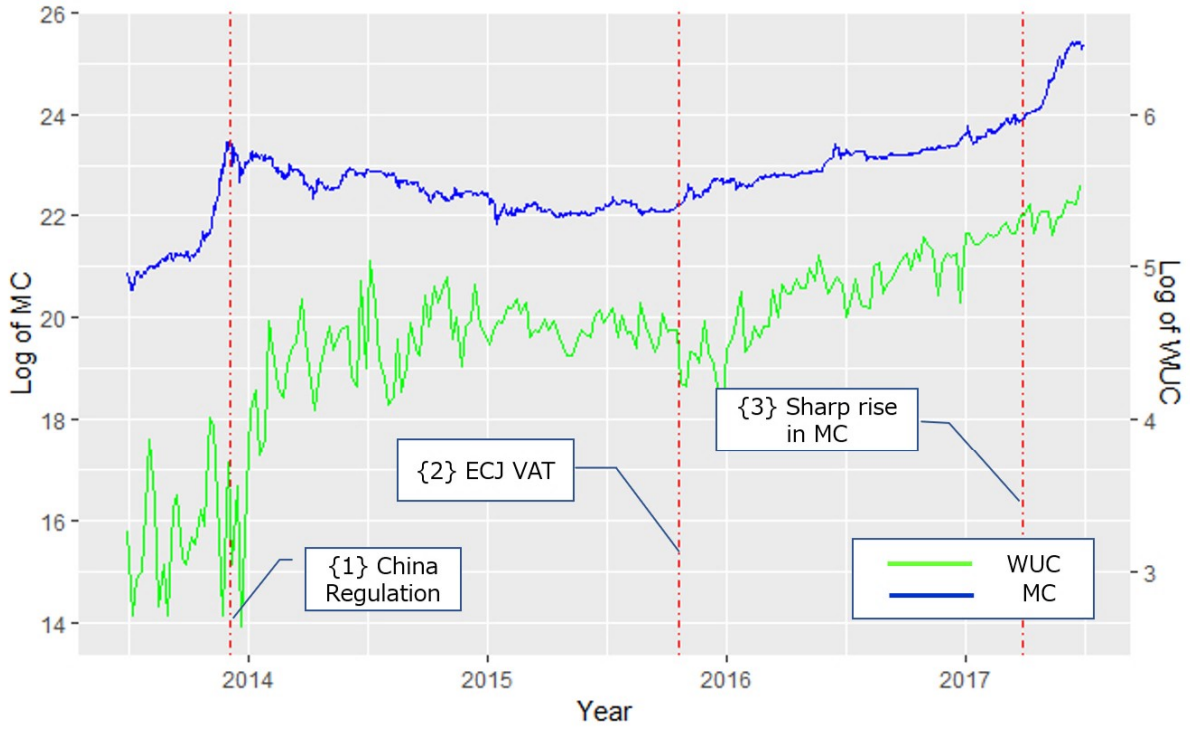

図 4.3　Time series of the sum of market capitalization (MC) and weekly unique contributors (WUC) of all cryptocurrency projects.

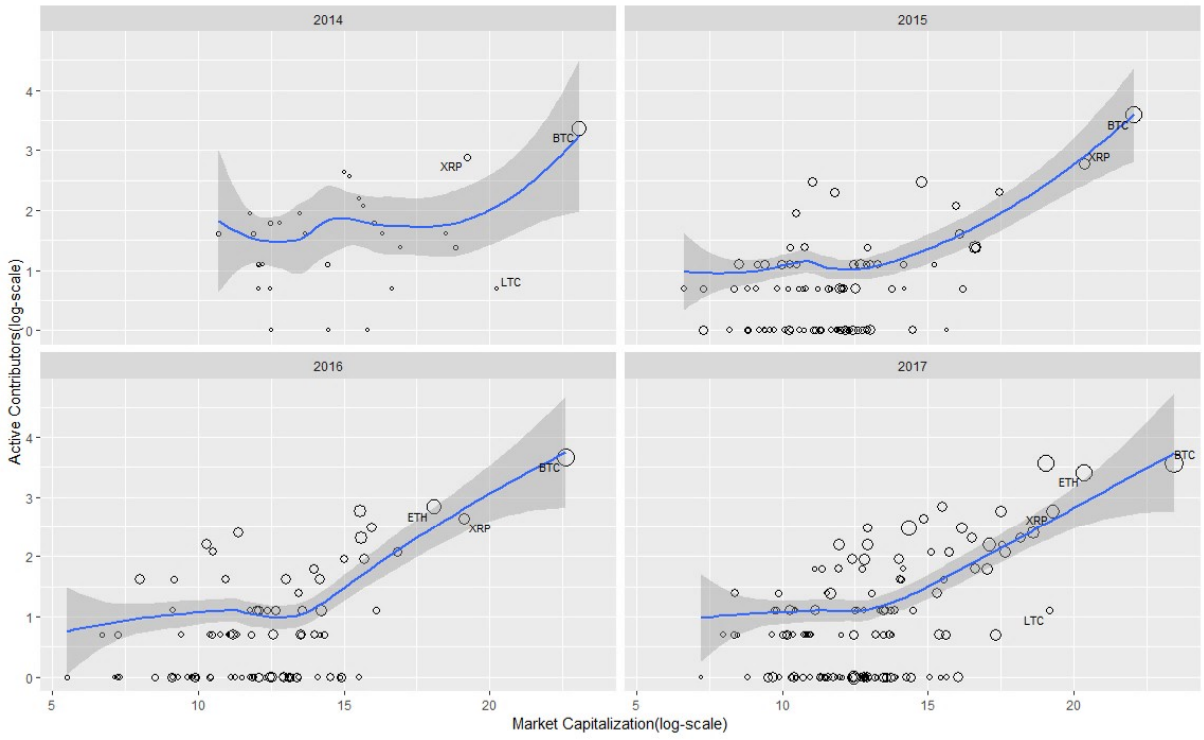

図 4.4　Relationship between active contributors and the market capitalization in log-scale with loss curve.

WUC の両方を自然対数変換し、正規分布に近づけた。

時系列分析に関係する方程式は以下のとおりである。

$$\begin{aligned} WUC_t &= c_1 + \delta_1 t + \phi_{1,1}^{(1)} MC_{t-1} + \phi_{2,1}^{(1)} WUC_{t-1} \\ &\quad + \cdots + \phi_{1,p}^{(1)} MC_{t-p} + \phi_{2,p}^{(1)} WUC_{t-p} + \varepsilon_{1,t} \end{aligned} \tag{4.1}$$

$$\begin{aligned} MC_t &= c_2 + \delta_2 t + \phi_{1,1}^{(2)} MC_{t-1} + \phi_{2,1}^{(2)} WUC_{t-1} \\ &\quad + \cdots + \phi_{1,p}^{(2)} MC_{t-p} + \phi_{2,p}^{(2)} WUC_{t-p} + \varepsilon_{2,t} \end{aligned} \tag{4.2}$$

$$\begin{aligned} WUC_t &= c_1' + \delta_1' t + \phi_{2,1}^{(3)} WUC_{t-1} \\ &\quad + \cdots + \phi_{2,p}^{(3)} WUC_{t-p} + \varepsilon_{1,t}' \end{aligned} \tag{4.3}$$

$$\begin{aligned} MC_t &= c_2' + \delta_2' t + \phi_{1,1}^{(4)} MC_{t-1} \\ &\quad + \cdots + \phi_{1,p}^{(4)} MC_{t-p} + \varepsilon_{2,t}' \end{aligned} \tag{4.4}$$

ここで、c_i は定数で、δ_i はトレンド項である。モデルの各パラメーター（δ および ϕ）は、最小二乗法 (OLS）によって推定可能である。

最初に、グレンジャー因果性検定 [Gra69] を実行して、MC から WUC への因果関係の存在を調べた。 [Gra69] によると、時系列 X が時系列 Y の予測に役立つ場合、時系列 X から Y にグレンジャー因果関係が存在すると言う。具体的には、Y の現在と過去の値のみに基づく将来の Y の予測と、Y と X の現在および過去の値に基づく将来の Y の予測を比較し、後者の平均二乗誤差（MSE）が小さい場合、X から Y のグレンジャー因果関係が存在する。一般的にグレンジャー因果関係は、因果関係の必要条件とみなされている。

グレンジャー因果性を検定するには定常時系列である必要がある。また両方の時系列の間に共和分の関係があってはならない。検定の結果、共和分の関係は存在しなかったが、定常性は確認できなかったため、一階差分により定常時系列を作成した。

次に、このモデルの残差の分散を計算した [Ham94]。最適なラグである $p = \{1, \ldots, 3\}$ は、AIC（赤池情報量基準）の最小値で決定した。ここで、SSR0 を式 4.1 の残差平方和、SSR1 を式 4.3 の残差平方和とする。式 4.5 を使用して F 統計の値を計算した。T はサンプルサイズである

$$Fstatistic = \frac{(\mathrm{SSR0} - \mathrm{SSR1})/2}{\mathrm{SSR1}/(T - 2p - 1)} \tag{4.5}$$

結果の F 統計量を F 分布の 95 % ポイントと比較すると、MC から WUC へのグレンジャー因果関係の仮説を統計的に検定できる。同じことを逆方向、式 4.2 と 4.4 を使用した WUC から MC への検定にも適用した。

表 4.2　Result of the Granger causality test

Case	Null signif.	Value	Result
1	$MC \longrightarrow WUC$	F statistic	5.276
		p value	0.0002^{***}
2	$WUC \longrightarrow MC$	F statistic	1.597
		p value	0.165

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1

表 4.2 に、グレンジャー因果性検定の結果を示す。 MC から WUC のケース 1 では、0.1 % 有意であった。ただし、逆方向のケース 2 では、有意ではなかった。したがって、MC から WUC へのグレンジャー因果関係が存在すると言える。逆に、WUC から MC へのグレンジャー因果関係は存在しなかった。

4.4.3　MC と WUC の時間ラグ

式 4.1 および 4.2 により、時間ラグの分析が可能となる。表 4.3 は、回帰分析の結果をまとめたものである。ケース 1 の場合、MC から WUC までの係数 $\phi_{1,2}^{(1)}$（2 か月前の MC）は、正の値で 1 % 有意であった。 これは、WUC の変化が MC の変化より 2 か月遅れる傾向があることを意味する。係数は、時価総額が 1% 増加すると、貢献者数が約 0.6% 増加することを示している。 WUC から MC までのケース 2 では、過去の MC の係数だけが有意であり過去の WUC 係数はいずれも有意ではなかった。 つまり、過去の WUC 値は将来の MC 値に影響を与えないことが確認できた。

4.5　本章のまとめ

上記の結果は次のように解釈できる。

- グレンジャー因果関係は因果関係が存在するための必要条件とみなされる。すなわち、表 4.2 から、グレンジャー因果性検定の結果は、1）時価総額から貢献者数への因果関係がある可能性があり、2）逆に貢献者数から時価総額への因果関係が存在しないことを示している。

- 回帰分析の結果（表 4.3 ケース 1)、時価総額が増加（減少）してから 2 か月後に貢献者の数が増加（減少）した。これは、プロジェクトの将来性（すなわち、時価総額）がプロジェクトへの参加を促すことを示唆していると考えられる。つまり時価総額の増加は、新しい貢献者の獲得につながると推測される。
- 逆に、過去の貢献者数と将来の時価総額の間には相関関係は確認できなかった（表 4.3 ケース 2)。時価総額を決定する要因は、その仮想通貨の人気や公共での関心（新聞やソーシャルメディアなど）に大きく影響されていると推測される。貢献者数の増加を、投資をおこなう一般的な人達が知ることはほとんど有り得ない。したがって、これは妥当な結果と考えられる。
- 回帰分析の結果は、グレンジャー因果性検定の結果と一致している。

本研究の結果は、時価総額と貢献者の数の間に真の因果関係が存在することを証明していない。時価総額と貢献者数の両方に影響を与える交絡因子が存在し、それが本当の原因である可能性がある。ただし、そのような交絡因子の存在を想定したとしても、そのような因子は仮想通貨の将来性を表すものと解釈できる。たとえば、 [GS15, PG17] らの研究は、仮想通貨の時価総額に対するソーシャルメディアの影響の存在を示している。ソーシャルメディアの書き込み内容は交絡因子である可能性があり時価総額と貢献者の参加の両方に影響を与えているかもしれない。しかしながらソーシャルメディアが時価総額に影響を与えるということであれば、それ自体が将来性の代理変数であり、本研究の重要な結論、つまり「プロジェクトの将来性はその貢献者の参加に影響を与える」を揺るがすことはない。

表 4.3　Result of regression analysis

Case 1 $MC \longrightarrow WUC$: $eq.(1)$			Case2 $WUC \longrightarrow MC$: $eq.(2)$		
Coefficient	Estimate	$P(>\|t\|)$	Coefficient	Estimate	$P(>\|t\|)$
$\phi_{1,1}^{(1)}$	-0.195	0.066^{+}	$\phi_{1,1}^{(2)}$	1.592	$8.55e-13^{***}$
$\phi_{2,1}^{(1)}$	0.292	0.077^{+}	$\phi_{2,1}^{(2)}$	-0.215	0.369
$\phi_{1,2}^{(1)}$	0.587	0.002^{**}	$\phi_{1,2}^{(2)}$	-1.093	$1.21e-4^{***}$
$\phi_{2,2}^{(1)}$	0.218	0.118	$\phi_{2,2}^{(2)}$	-0.260	0.203
$\phi_{1,3}^{(1)}$	-0.161	0.262	$\phi_{1,3}^{(2)}$	0.575	0.009^{**}
$\phi_{2,3}^{(1)}$	0.063	0.59	$\phi_{2,3}^{(2)}$	0.050	0.774
C_1	-3.331	0.005^{**}	C_2	-0.036	0.982
δ_1	0.004	0.209	δ_2	0.013	0.011^{*}

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1

第 5 章

構造トピックモデルを用いた貢献ガイドラインと貢献量との関連性の研究

本章では、GitHub の標準ファイルである貢献ガイドラインについて分析をおこない、貢献ガイドラインの記載内容が貢献者数と関連があるか確認する。また、貢献ガイドラインの記載内容を考察し、本来記載すべき内容を提言した。構造トピックモデルを用い記載内容をトピックに分解した後、各トピックと貢献者数との関連性を分析する。

5.1 序論

ソフトウェア開発においてガイドラインは根幹をなす。ガイドラインが不備であったり、ガイドラインに従わないでコーディングしたりすると規則性が失われ不具合を生じやすくなり、メンテナンスの負荷が高くなる。ガイドラインの効果を検証し、品質を向上させることは、OSS 開発プロジェクトの生産性を高めることにつながる。

本研究では、GitHub の標準ファイルである貢献ガイドラインに着目する。 貢献ガイドラインは GitHub の標準ファイルであり（contributing.md/contributing.txt [*1]）、各開発プロジェクトで用意することが推奨されている。ユーザーは貢献ガイドラインの記載内容に従い貢献活動をすることが期待される。GitHub は、ユーザーが不具合報告や Pull Request（修正の反映要求）をする際に、ガイドラインへのリンクを表示し、参照を促して

[*1] https://github.com/blog/1184-contributing-guidelines (accessed 2020-07-23)

いる。貢献ガイドラインには、コーディング規約やパッチリリース手順などの項目の記載が推奨されているが、各項目は明確に定義されておらず、プロジェクトごとに異なっているのが現状である。それは、ユーザーを混乱させるだけでなく、期待した記載がない為にユーザーは貢献をとりやめる可能性もある。また、貢献ガイドラインの有効性についても十分に分析されてこなかった。貢献ガイドラインの存在がプロジェクト生存率を高める研究 [CP16] は存在するが、貢献ガイドラインの記載内容まで踏み込んだ研究は存在しない。

本章では、構造トピックモデル (structural topic model: 以下 STM) を用い、貢献ガイドラインをトピックに分解し、記載内容を分析する。その上で、貢献者数と貢献ガイドラインの内容との関連性を分析し、貢献ガイドラインの内容がプロジェクト成功の指標である貢献者数に及ぼす影響を検証するとともに、貢献ガイドラインで重要と考えられる記載内容を考察する。

この章の構成は以下のとおりである。

最初に 5.2 章で STM について説明する。次に 5.3 章において、分析で使用されるデータセットについて説明し、その概観を示す。 5.4 章では、データ検証の手順およびその結果を報告し、最後に 5.5 章で結果をまとめる。

5.2 構造トピックモデル

本研究では貢献者数・貢献ガイドラインに付随する情報と記載内容の関連性を分析する為に STM を用いた。STM はトピックモデルの一種である。トピックモデルは、文書と単語の間の共起行列（bag-of-words）を、文書とトピック、トピックと単語、の 2 種類の行列に分解する手法である。文書ごとのトピック、トピックごとの単語、それぞれの出現確率を出力するため定量的に解釈し易い。STM は、トピック分解に加え、文書に付随する属性情報・数値情報（共変量）によるトピックの出現確率への影響を測定できる特徴を持つ [RSTA13]。共変量の効果は 2 種類存在する。文書ごとのトピックの出現確率に影響を与える共変量 (Topic Prevalence) とトピックごとの単語の出現確率に影響を与える共変量（Topic Content) である。

それらを考慮し、STM では、各パラメータの同時分布を以下のように展開する。

$$p(w, z, \theta, \beta) = p(w|z, \beta)p(z|\theta)p(\theta|\mu)p(\mu|\gamma X, \Sigma)p(\beta|m, k, Y) \tag{5.1}$$

θ : 各文書に出現するトピックの確率

β：各トピックに出現する単語の確率

z：トピック

w：単語

X：トピックの出現確率に影響を与える共変量（Topic Prevalence)

Y：単語の出現確率に影響を与える共変量（Topic Content)

$p(\mu|\gamma X, \Sigma)$：は平均が γX, 分散が Σ の多変量対数正規分布

m：ベースライン単語分布、k:トピックにおける偏差

今回は共変量によるトピックの出現確率の変化に着目するので Y は使用せず、X のみ用いて分析を実施した。

STM は事前分布 ($p(\mu|\gamma X, \Sigma)$) を多変量対数正規分布にして、共変量（X) がトピックの出現確率の平均 γX に影響を与えるモデルである。近年は、ソーシャルメディアのコンテンツ解析にも利用され始めている [CF19, FPSF19, BNL20]。

5.3 データセット

5.3.1 データセットの準備

貢献ガイドラインを含むレポジトリ (GitHub におけるプロジェクトやファイルの管理スペース) の割合は極めて低く、一定数を確保するためには大量のレポジトリを探索する必要がある。GitHub のレポジトリからデータを探索する方法は、GitHub が提供している application interface(API) を利用することが一般的であるが、トラフィック量に上限があり大量の探索には適さない。そこで本研究では、GitHub アーカイブであるアイオワ州立大学が提供する BOA [DNRN13] のデータセット (BOA SPEC2015 Data) を使用して、貢献ガイドラインを含むレポジトリを抽出した。 BOA は、Hadoop テクノロジーを採用した GitHub レポジトリの巨大なアーカイブであり、短時間に対象レポジトリとその属性の取得が可能である。該当アーカイブには 7,830,023 件のレポジトリが含まれているが、探索の結果、貢献ガイドラインが存在したのは 620 レポジトリであった。

BOA のデータセットには貢献ガイドラインの URL だけが含まれており、ガイドライン自体は含まれていない。取得した URL にアクセスして貢献ガイドラインをダウンロードした。レポジトリやファイルが削除されているケースもあり、ダウンロードできたガイドラインは 459 であった（2017 年 4 月時点)。次に GitHub API を使用して、貢献ガイド

ラインが用意されているレポジトリから貢献者の数を取得した。

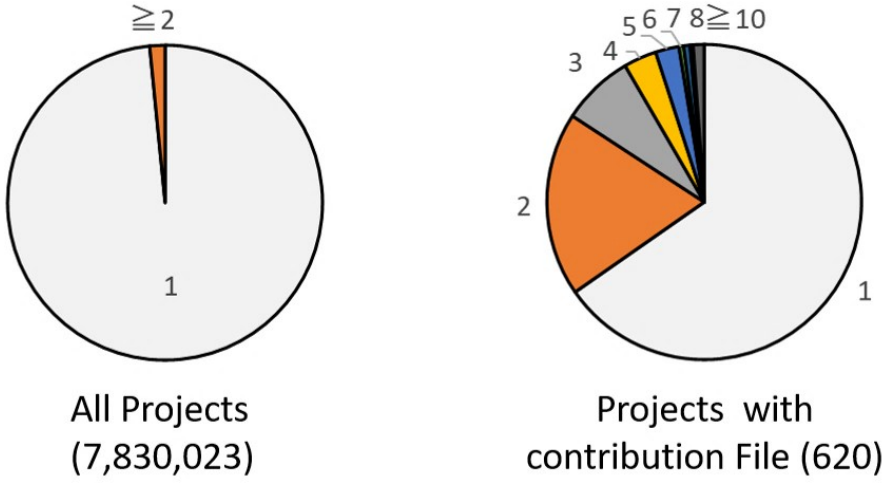

図 5.1 The ratio of contributors to GitHub projects. The 98.4% of projects have only one contributor, i.e. owner only (left). The 65.3% of repositories with contributing.md have one contributor (right). (Source: BOA SEP2015 Data)

5.3.2 データセットの概観

GitHub は、ソフトウェア共同開発プロジェクトのサポートを目的としたサービスであるが、無料のファイルストレージやバージョン管理システムの学習など、実際には貢献者を必要としない多岐の用途に使われている。GitHub のレポジトリが、開発プロジェクトかそれ以外の用途で使われているかを識別することは容易ではなく、従来は人手により分類していた [PTT+19]。貢献ガイドラインは開発プロジェクト以外には必要ではない、それゆえ貢献ガイドラインがリポジトリに存在すれば、それは共同開発プロジェクト用途であるとみなすことができる。

貢献ガイドラインを含むレポジトリ数は 620 で全体の 0.008% と極めて低い。同じ GitHub の標準ファイルである readme.md は、183,956 レポジトリに存在したので約 300 分の 1 である。 readme.md はファイルストレージなどの多岐の用途でも必要であり [PTT+19]、その差があらわれたものと考えられる。GitHub アーカイブの全 7,830,023 レポジトリのうち、貢献者が 1 名（所有者だけ) の割合は 98.4% （図 5.1 左）であった。

それらの多くのレポジトリは共同開発以外の目的での利用と推測される。一方、貢献ガイドラインを含むレポジトリ (620) では貢献者 1 名の割合は 65.3%（図 5.1 右）であった。貢献ガイドラインを用意していることから外部からの貢献を期待している開発プロジェクトと考えられるが、その場合でも、3 分の 2 は貢献者が獲得できていない。

外部からの貢献を期待する場合、貢献ガイドラインにはどのような記載をすべきだろうか。GitHub の管理ページでは「良い例」として puppet プロジェクトのガイドラインが示されている [*2]。そこには、不具合報告、コードの変更方法、パッチのリリース手順などの項目が含まれている。一方、必要と思われる環境構築やコーディング規約については記載がない。

一般的に、開発プロジェクト設立時にはこの例に準拠にして貢献ガイドラインを作成すると思われるが、実際はどうなのであろうか？この事を確認するため、予備調査として GitHub の上位 6 プロジェクト[*3]の貢献ガイドラインを調査した。図 5.2 は、上位 6 プロジェクトの貢献ガイドラインに含まれている単語をワードクラウド化したものである。ワードクラウドでは出現頻度が高い単語ほどフォントサイズは大きくなり中心に近づく。プロジェクトごとに大きく外観が異なっているのがわかる。前述した puppet は左上であるが、各プロジェクトは単純にそれに準じているようには見えない。つまり各プロジェクトは自プロジェクトの必要性やコンセプトに応じた貢献ガイドラインを作成しているものと推測される。

5.4 データ解析

5.4.1 共起行列の作成

前処理として分析に適さない貢献ガイドラインを以下の手順で除外し、抽出した単語を使って文書と単語の間の共起行列を作成した。

1. 貢献ガイドラインが極めて短い場合、実際のガイド内容が記述されているサイトへのリンクなど貢献内容の記載がない。同じ GitHub の標準ファイルである README の研究においては、一定サイズ以下のファイルを除外する対応をしてい

[*2] https://github.com/puppetlabs/puppet/blob/master/CONTRIBUTING.md（accessed 2020-07-03）

[*3] https://github-ranking.com/ （accessed 2017-04-22）

図 5.2　Word clouds from contributing.md in top GitHub projects. (This data was obtained on 2017-04-22)

る [PTT+19]。本研究においても同様に 100 語に満たない貢献ガイドラインを分析対象から除外した。その結果 407 ファイルが残った。

2. 同じ組織に属するプロジェクトで同じガイドラインを使い回しているケースが存在した。個別に調査し、代表的なプロジェクト（貢献者数が最も多い）のみを残し他のプロジェクトは除外した。その結果 245 ファイルとなった。
3. 各ガイドラインから単語を抽出した。対象は名詞、動詞、形容詞、副詞とし、各単語の語形の変化を取り除き同一の単語表現に変換するするステミング処理をおこなった。
4. TF-IDF により各単語の重みづけを調整した。TF-IDF は、文書中に含まれる単語の重要度を評価する手法の 1 つであり、主に情報検索やトピック分析などの分野で用いられる。Moh らは、TF-IDF により文書集合固有のストップワードの重みが軽くなり、より正確なクラスタリングが可能となるとしている [MB12]。 TF-IDF は、TF（term frequency）と IDF（inverse document frequency）を掛け合わせた値である。TF とは、単語の文書内の出現頻度である。ある文書中に出現する割合が高い単語ほど重要である可能性が高い。IDF は、ある単語が出現する文書数の逆数と関連する指標である。多くの文書中に出現する単語は、特徴語とは言い難いため IDF の値は小さくなる。

   TF-IDF は、TF（Term Frequency、単語の出現頻度）と IDF（Inverse Document Frequency、逆文書頻度）の二つの指標に基づいて計算される。

   TF は、単語の文書内の出現頻度である。ある文書中に出現する割合が高い単語ほど重要であると考えられる。

$$tf(t,d) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}} \tag{5.2}$$

   tf(t,d)：文書 d 内のある単語 t の TF 値

   $n_{t,d}$：ある単語 t の文書 d 内での出現回数

   $\sum_{s \in d} n_{s,d}$：文書 d 内のすべての単語の出現回数の和

   IDF は、ある単語が出現する文書数の逆数と関連する指標である。多くの文書中に

出現する単語は、特徴語とは言い難い。

$$idf(t) = log\frac{N}{df(t)} \tag{5.3}$$

idf(t)：ある単語 t の IDF 値

N：全文書数

df(t)：ある単語 t が出現する文書の数

TF-IDF 処理前と処理後の重要語を表 5.1 に示す。処理前では、make, use, file, sure, new など極めて汎用的で文書の特徴を表現し難い単語が上位に含まれている。一方、処理後にはそれらの単語は排されている。

5.4.2 トピック分解

次に上記で得られた共起行列を STM によりトピックに分解した。トピックの出現確率との関連性を調べるため、貢献者数と貢献ガイドラインに付随する属性情報を共変量 X とした。取得可能な属性情報は、貢献ガイドライン作成からの日数、ガイドライン更新の有無、ガイドラインの単語数である。貢献者数については正規分布の適合性の確認をおこない、対数変換した。

この時トピック数は下記手順で選択した。

1. 事前にいくつかの貢献ガイドラインを抽出して人手により記載内容を調べ、凡のトピック数を推測した。その結果 7 つのトピック (Question, Bug-report, New-feature, Change-rules, Workflow, Coding-guide, License) を洗い出した。
2. Robert らはトピック数決定の指標として Semantic Coherence(文書内での単語共起の指標), Residuals(残差), Held-Out Likelihood(尤度), Lower Bound(変分法を使った解析の下界) を提案している [RST19]。前述の手順でトピック数を 7 前後と推定したので、その近傍のトピック数 (4-12) について、それぞれの値を計算した (図 5.3)。
3. Semantic Coherence はトピック数 6, 7 が一番良い値となっている。トピック数 7 の方が 6 より他の指標が良い（Held-Out Likelihood と Lower Bound が高く、Residual が少ない)。以上から 7 をトピック数として選択した。

表 5.1　TOP 20 important terms

TF-IDF 処理前				**TF-IDF 処理後**		
rank	**term**	**sum**	**mean**	**term**	**sum**	**mean**
1	change	1713	4.21	commit	3403	10.98
2	code	1323	3.25	license	3380	10.90
3	make	1302	3.20	git	3268	10.54
4	commit	1151	2.83	run	3183	10.27
5	test	1108	2.72	issue	3133	10.11
6	issue	1087	2.67	change	3007	9.70
7	use	1064	2.61	style	2953	9.53
8	submit	1010	2.48	feature	2929	9.45
9	project	953	2.34	contribution	2913	9.40
10	contribution	925	2.27	add	2822	9.10
11	file	888	2.18	report	2762	8.91
12	repository	858	2.11	branch	2735	8.82
13	contribute	853	2.10	ticket	2722	8.78
14	work	843	2.07	submit	2635	8.50
15	branch	795	1.95	repository	2545	8.21
16	follow	752	1.85	test	2506	8.08
17	sure	695	1.71	sign	2502	8.07
18	bug	659	1.62	bug	2490	8.03
19	create	645	1.58	project	2348	7.57
20	new	556	1.37	send	2343	7.56

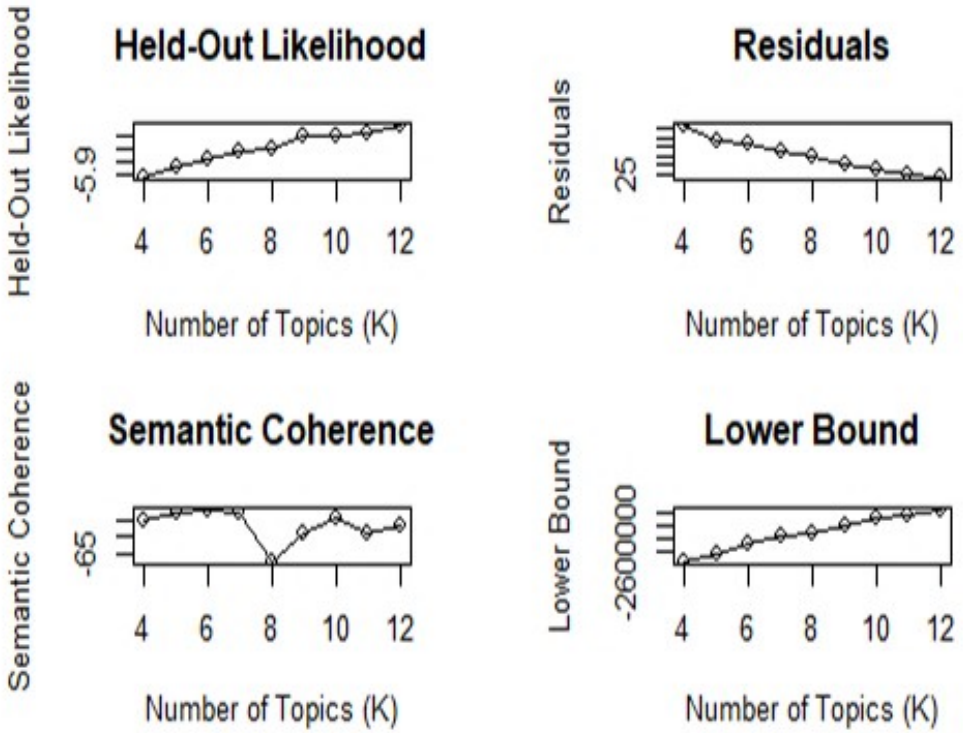

図 5.3　Diagnostic values by number of topics

表 5.2　Examples of topic sentences

No.	Topic Titles	Sentences
1	Build Environment	You can then import the 4 plugin projects into Eclipse.
2	Package License	I have the right to submit it under the open source license.
3	Git Clone	If you cloned a while ago, get the latest changes from upstream.
4	Issue Report	If you want to report an issue, please make sure to be concise.
5	New Feature	Encourage you to contribute to the open source projects by implementing new features.
6	Format Code	Follow existing conventions and style in order to keep the code as readable as possible.
7	Git Update	Commit your changes into that branch.

各トピックと上位 10 単語の図 5.4 に示す。横軸はトピックに含まれる単語の出現確率 (β) である。各トピックのタイトルは上位の単語を参考にして命名した。それぞれのトピックの代表的な文を表 5.2 に記載する。Build Environment は環境構築に関するトピックである。これについては手作業での抽出では全く見落としていた。また手作業では Git の処理関連（コードの変更方法、パッチのリリース手順など）の単語をまとめて Change-rule としていたが、STM では貢献者の環境準備 (Git Clone: Git クローン操作) と修正内容反映 (Git Update: Git 更新操作) に分離されている。Format Code はコーディング規約である。このトピックは上位単語の出現確率が高く最も明確に分離されている。Issue Report、New Feature は不具合報告、新規機能とした。Topic 2 を License Package と命名したのは、両方に関連した単語が混在するためである。トピック数を増して両者が分離できるか試してみたが、Package は分離するが License が複数のトピックに分散し、却って全体の品質が低下した。そのため、トピック数は 7 のままとした。

5.4.3　トピックの評価

STM は各文書（分析の対象とした 245 の貢献ガイドライン）について、各トピックの出現確率を計算する。表 5.3 に 245 文書における各トピックの出現確率の平均を示す。Topic 4 が若干大きく、Topic 6 が若干小さいが大きなばらつきはみられない。

図 5.5 は、各トピックの品質について示したものである。横軸が Semantic Coherence、縦軸が Exclusivity であり、それぞれ値が大きいほど性能が良い。Semantic Coherence とは、各トピックで上位確率の単語が文書内で共起する傾向の指標であり、Mimno らは以

図 5.4　Highest word probabilities for each topic

表 5.3　The rate of topics (N=245)

Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**	**Topic6**	**Topic7**
Buid environment (環境構築)	License Package (ライセンス・パッケージ)	Git Clone (Git クローン操作)	Isuue Report (不具合報告)	New Feature (新規機能)	Format Code (コーディング規約)	Git Update (Git 更新操作)
13.4	13.4	14.4	18.8	15.7	9.1	15.2

下の式での計算を提案している [MWT+11]。

$$SemanticCoherence(W) = \sum_{w1,w2 \in W} log \frac{D(w1,w2)+1}{D(w2)} \tag{5.4}$$

ここで、W は各トピックでの上位の単語セット (10 単語) であり、w1、w2 はそれに含まれる単語 ($w1 \neq w2$)、D は各単語の出現確率である。D(w1,w2) は同じ文書で w1、w2 が同時に出現する確率。分子に 1 を加えているのは対数の中身がゼロにならないようにするためである。本研究ではこの式を用いて上位確率の単語が文書内で共起する傾向を分析した。

あるトピックで出現確率が高い単語が他のトピックで出現確率が低い傾向を表す指標が Exclusivity（排他）である。あるトピックの上位単語の出現確率を、全トピックでのその単語の出現確率の合計で割ったものである。本研究では出現確率の項 (Frequency) を加えた FREX という指標を使用した [BA12, RSA16]。これはトピック内で出現確率が高い単語ほど FREX への貢献度が高くなるように補正した指標で、以下の式で求められる。

$$FREX = (\frac{\omega}{ECDF\left(\frac{\beta}{\sum_{j=1}^{K} \beta_j}\right)} + \frac{1-\omega}{ECDF(\beta)})^{-1} \tag{5.5}$$

ω : Exclusive と Frequency の割合、ここでは 0.5

β : 各トピック内で単語が出現する確率

ECDF (Empirical Cumulative Distribution Function) : 経験的累積関数。値の間隔が均等に近づくように補正している。

図 5.5 によると、Topic 6(コーディング規約)、Topic 1(環境構築)、Topic 3(Git クローン操作) はどちらの値も高い。Topic 2(ライセンス・パッケージ)、Topic 5(新規機能) は Semantic Coherence が低い。Topic 4(不具合報告)、Topic 5(新規機能) は Exclusivity が低い。図 5.4 を見ると、Topic 6,1,3 には出現確率が相対的に高く、且つ、他のトピックには余り含まれない特徴的な単語が存在しており、トピックとして明確である。一方、Topic 5 には commit や report など他のトピックでも上位に位置する単語が含まれており、且つ、その出現確率は高くない。他の単語と共起しない残りの単語を集めたトピックと考えられる。

図 5.6 は各文書のトピック別の γ 値のヒストグラムである。全体的に各トピックが均質

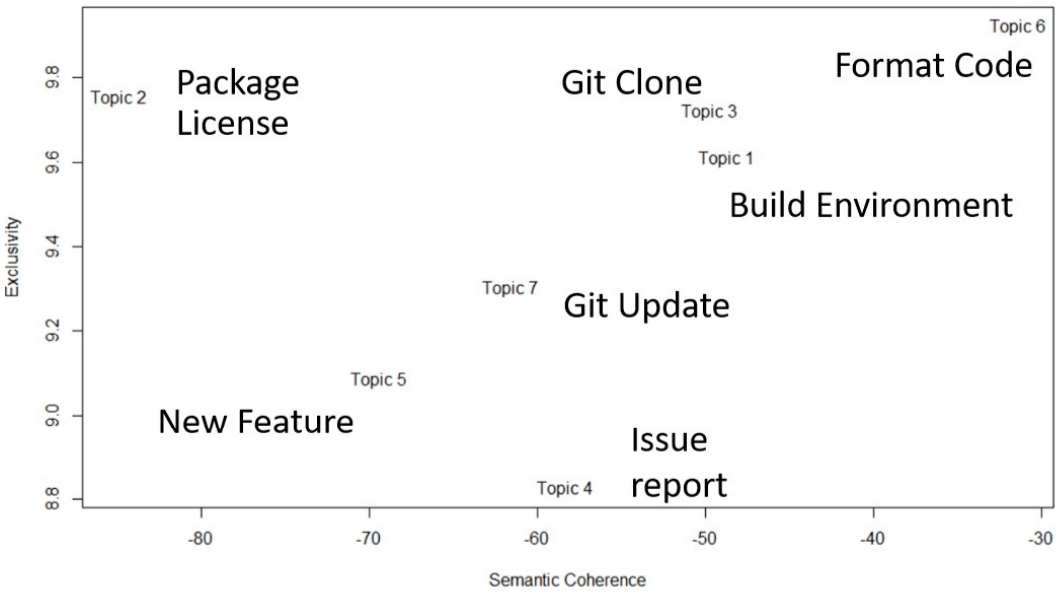

図 5.5　Topic quality

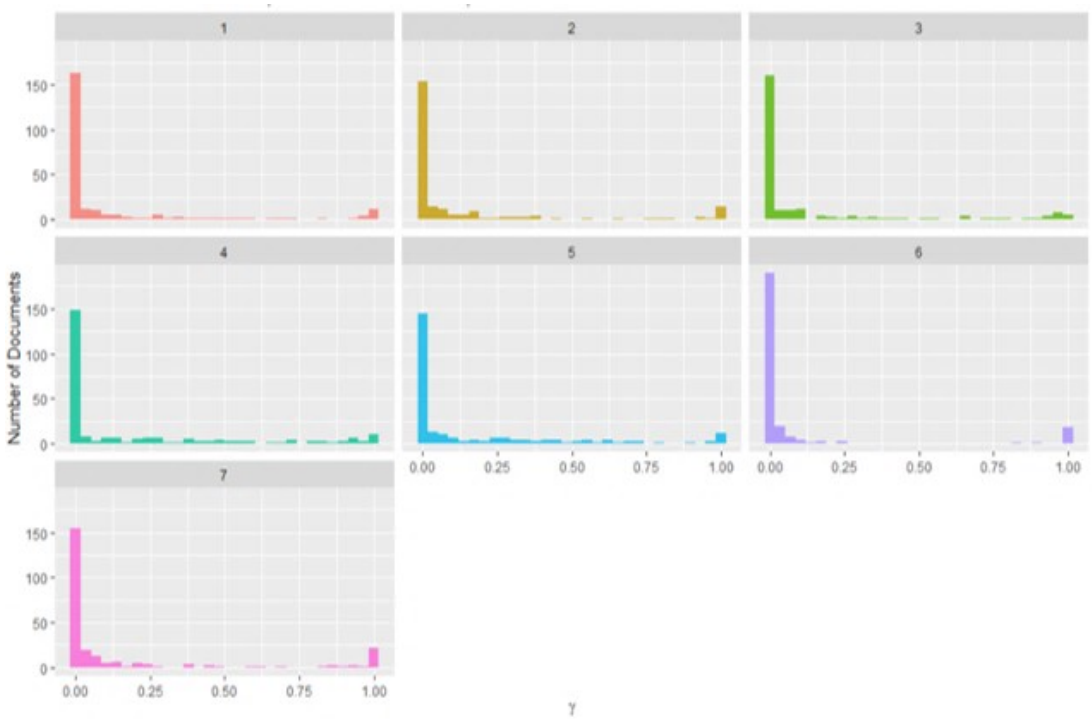

図 5.6　Distribution of document probabilities for each topic

であるが、Topic 6(Format Code) を全く含んでいない文書の割合が若干高い。

5.4.4　各共変量とトピックの関連性

貢献者数は Topic 4 において増加傾向がみられる、また Topic 7 において減少傾向がみられる（図 5.7)。STM により分析した貢献者数と各トピックの出現確率との関連性を表 5.4 に示す。貢献者数 (Contributors) は Topic 4 が 1% 水準で有意で正の相関がみられた。また Topic 7 は 5% 水準で有意であり負の相関が存在した。Topic 7 は切片が 0.24 と高く、0.1% 水準で有意であった。

表 5.5 は貢献ガイドラインの属性と各トピックの出現確率との関連性を示したものである。単語数 (Length) は Topic 4 が正の相関で 5% 水準で有意、Topic 6, 7 が負の相関でそれぞれ、1%、5% で有意だった。ガイドライン作成からの日数 (Days) は特に顕著な傾向はみられなかった。貢献ガイドライン更新 (Modify) は Topic 1 が 1% 水準有意で正の相関、逆に Topic 7 は 1% 水準有意で負の相関が存在した。Topic 6, 7 は切片の値が他のトピックと比べ極めて高く（両者を合わせて 6 割以上)、それぞれ 0.1% 水準、1% 水準で有意であった。

表 5.4　Prediction of topic probability using contributors (N=245)

	Topic 1				Topic 2				Topic 3				Topic 4			
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	9.31E-02	4.14E-02	2.248	0.026*	1.42E-01	4.05E-02	3.497	0.001***	1.45E-01	4.16E-02	3.483	0.001***	7.00E-02	4.33E-02	1.618	0.107
Contributors	4.02E-02	3.43E-02	1.17	0.243	-7.25E-04	3.26E-02	-0.022	0.982	-5.20E-03	3.34E-02	-0.156	0.876	9.32E-02	3.48E-02	2.676	0.008**

	Topic 5				Topic 6				Topic 7			
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	1.43E-01	4.02E-02	3.555	0.000***	1.60E-01	3.79E-02	4.232	0.000***	2.46E-01	4.29E-02	5.748	0.000***
Contributors	6.54E-03	3.24E-02	0.202	0.840	-5.08E-02	2.98E-02	-1.706	0.089.	-8.19E-02	3.38E-02	-2.427	0.016*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

表 5.5　Prediction of topic probability using contributing.md attributes (N=245)

	Topic 1				Topic 2				Topic 3				Topic 4			
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-2.62E-02	9.75E-02	-0.269	0.789	1.30E-01	9.06E-02	1.431	0.154	1.83E-01	9.62E-02	1.9	0.059.	7.30E-03	1.05E-01	0.07	0.945
Modify	1.24E-01	4.19E-02	2.963	0.003**	-9.54E-04	4.38E-02	-0.022	0.983	-7.07E-03	4.39E-02	-0.161	0.872	7.10E-02	4.57E-02	1.554	0.122
Length	9.40E-07	2.74E-05	0.034	0.973	3.38E-05	2.86E-05	1.183	0.238	1.91E-06	2.47E-05	0.077	0.938	7.10E-05	3.26E-05	2.178	0.030*
Days	5.07E-05	6.09E-05	0.833	0.406	-8.75E-06	5.45E-05	-0.161	0.873	-2.74E-05	5.78E-05	-0.474	0.636	4.49E-05	6.36E-05	0.705	0.481

	Topic 5				Topic 6				Topic 7			
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	7.15E-02	1.03E-01	0.696	0.487	3.08E-01	8.49E-02	3.63	0.000***	3.26E-01	9.85E-02	3.309	0.001**
Modify	1.16E-02	4.42E-02	0.261	0.794	-6.43E-02	3.99E-02	-1.612	0.108	-1.35E-01	4.57E-02	-2.947	0.004**
Length	1.15E-05	2.85E-05	0.405	0.686	-6.66E-05	2.34E-05	-2.847	0.005**	-5.25E-05	2.61E-05	-2.016	0.045*
Days	4.27E-05	6.24E-05	0.683	0.495	-7.07E-05	5.13E-05	-1.379	0.169	-3.04E-05	5.96E-05	-0.51	0.610

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

表 5.6 に更新されていない貢献ガイドラインと 1 度以上更新された貢献ガイドラインのトピックの出現確率の分布を示す。Topic 1 と Topic 7 は両者の間で差がみられた（5% 水準で有意）。貢献ガイドラインが未更新の場合 (Modify=N)、Topic 7 は平均が 0.256 と高く、Topic 1 は平均が 0.050 と極めて低い。未更新の貢献ガイドラインには、Topic 1 の記述はほぼ無いと言える。更新されている場合 (Modify=Y)、各トピックの出現確率に大きな差は見られない。

5.4.5 STM による結果の解釈

図 5.4 で命名した各トピックのタイトルが妥当であると仮定した上で、図 5.5, 表 5.3、表 5.4、表 5.5、表 5.6 に示した STM による定量的な分析結果を解釈すると以下のようになる。

1. 全体のトピックの出現確率は、不具合報告 (Issue Report) が若干大きく、コーディング規約 (Format Code) が若干小さい (表 5.3)。
2. 環境構築 (Build Environment)、コーディング規約 (Format Code)、 Git クローン操作 (Git Clone) は各指標が良好でありトピックとして明確である。新規機能 (New Feature) は残りの単語を集めたようなトピックであり明確ではない (図 5.5)。
3. 貢献者数と不具合報告 (Issue Report) トピックの出現確率との間には正の相関、Git 更新操作 (Git Update) との間には負の相関が存在する。ただし、Git 更新操作の切片の出現確率は高い (表 5.4)。
4. 不具合報告 (Issue Report) は単語数が多くなるとトピックの出現確率が高くなる。コーディング規約 (Format Code) と Git 更新操作 (Git Update) は逆に減少する。ただし、両トピックとも切片の出現確率は高い (表 5.5)。
5. 環境構築 (Build Environment) は未更新ガイドラインには、記載がほとんどない (表 5.6)。更新されたガイドラインには他のトピックと同等の出現確率があるので、更新時に追記されたと考えられる。
6. Git 更新操作 (Git Update) は未更新ガイドラインでのトピックの出現確率が高い (表 5.6)。更新されたガイドラインでは、他のトピックと同等の出現確率となる。
7. ガイドラインの各トピックの出現確率とガイドライン作成からの日数 (Days) との間に関連性は見られなかった (表 5.5)。

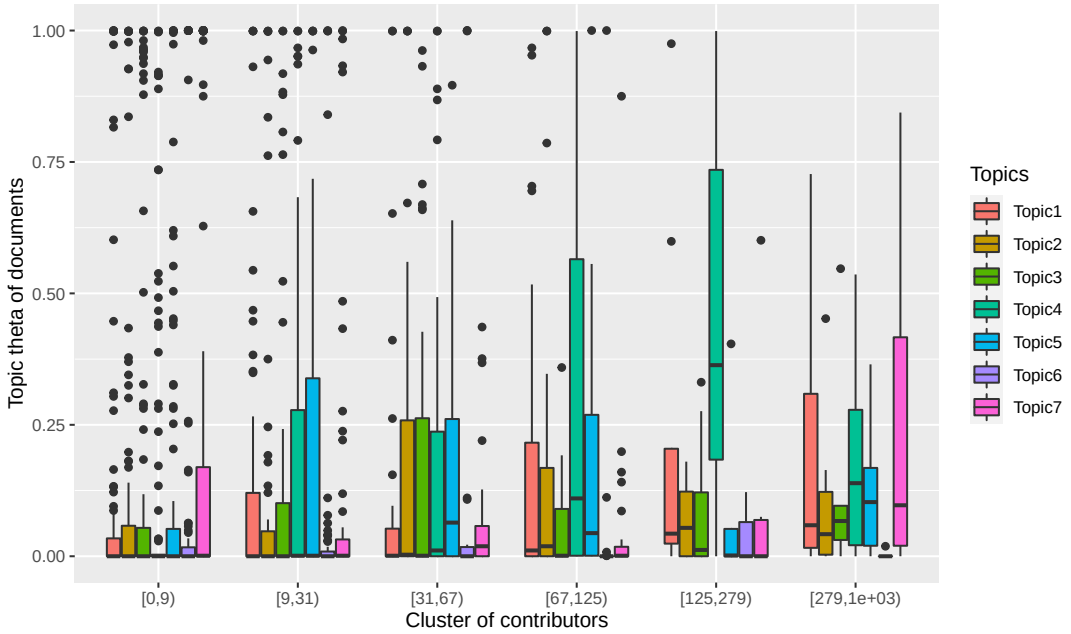

図 5.7　Topic theta of documents for contributor clusters

表 5.6　Effects on modification of the contributing.md for each topic

	Modify=N (N=94)			Modify=Y (N=151)		
	mean	2.5% CI	97.5% CI	mean	2.5% CI	97.5% CI
Topic1*	0.050	-0.006	0.110	0.174	0.117	0.232
Topic2	0.132	0.071	0.192	0.132	0.070	0.192
Topic3	0.144	0.081	0.210	0.136	0.077	0.196
Topic4	0.107	0.039	0.177	0.177	0.113	0.244
Topic5	0.141	0.075	0.207	0.152	0.091	0.215
Topic6	0.171	0.113	0.229	0.108	0.053	0.161
Topic7*	0.256	0.192	0.321	0.121	0.063	0.181

不具合報告手順が明瞭に書かれていると、ユーザーは安心してプロジェクトに参加しやすくなり、貢献者が増加することが考えられる。また Git 更新操作ばかり書いてあり、他の項目が書いていないようなガイドラインをユーザーは敬遠することもありうる。上記項目 3 は、そのような考察の一根拠となりうると考えている。また、貢献者が多くなると不具合報告の仕方も各自ばらばらになりがちである。それを防ぐために不具合報告の仕方を明確に書く必要がでてくる。プロジェクトにおける開発工程の初期では開発を重視し、終盤ではメンテナンスを重視するのが自然である。貢献者数は徐々に増加すると考えられるので、開発段階で重要な Git 更新操作 (Git Update) からメンテナンスの段階で重要な不具合報告 (Issue Report) にトピックが遷移していることが想定される。項目 4 は、そのような理解が可能である。

環境構築 (Build Environment) は「良い例」として示されているガイドラインには項目が存在しない。しかしながら実際の貢献ガイドラインには主要なトピックの 1 つとして存在しており、更新後のガイドラインに出現確率が増えている (項目 5) 事からも重要な役割をはたしている事が推定され、貢献ガイドラインへの記載が推奨される。同様に主要なトピックである「コーディング規約」、「Git 操作」、「不具合報告」も貢献ガイドラインへの記載が推奨される。一方、商用ソフトウェア開発で重視されているコーディング規約 (Format Code) はトピックの出現確率があまり高くない。OSS の開発はボランティアが中心で、自由を求める傾向がある。プロジェクトの管理者は、この規定を厳格に規定すると抵抗される、もしくは参加に躊躇されると考え、記載を控えた結果かもしれない。

5.5 本章のまとめ

GitHub には貢献ガイドライン以外にも標準ファイルが用意されており、どのファイルに何を記載すべきかユーザーを混乱させている。Prana らは、一番多く用意されている Readme.md について分析し、その記載内容を 8 つのカテゴリーに分類した。その項目として不具合報告、ライセンス、貢献など、貢献ガイドラインと重複した記載内容が含まれている [PTT+19]。しかしながら、開発者が参照すべき項目は貢献ガイドラインへの記載が推奨されており、それらは貢献ガイドラインに記載することが妥当であろう。本研究では貢献ガイドラインを対象に分析をおこない、以下を明らかにした。

- 貢献者数と「不具合報告」のトピックの出現確率の間には正の相関、「Git 更新操

作」との間には負の相関が存在する。

- 「環境構築」、「コーディング規約」、「Git 操作」、「不具合報告」は、貢献ガイドラインへの記載が推奨される。

第 6 章

総括と今後の研究展望

OSS 開発プロジェクトの体制は時代と共に変わっていった。そして研究テーマもそれに追従して「モチベーション」から「貢献者獲得」へと変化していった。

当初の OSS、Linux や Apache プロジェクトの開発体制は、階層的な組織構造であり、各メンバーの役割と責任が決められていて固定メンバーが中心であった。そうした状況下での研究テーマは、プログラマーが OSS に関わる「モチベーション」に関する分析が中心であり、1 つのプロジェクトを詳細に調査するものが多数を占めた。ただし、Linux や Apache 関連のプロジェクトを含む論文数は 2003 年にピークに達し、その後急激に低下した。

その理由は、GitHub が OSS の開発環境として人気を博したことである。GitHub には、使いやすい共同ソフトウェア開発ツールが備わっており、誰でも平等にプロジェクトに参加できる。GitHub のサービスは無償で提供されており、容易に OSS 開発プロジェクトを始めることができる。しかし、簡単にプロジェクトが始められる状況になると、貢献者の奪い合いが激化する。折角、斬新なアイデアをもってプロジェクトを開始したとしても、貢献者が集まらず失敗に終わることも多々ある。そうした状況下では「貢献者獲得」に関心が高まり、その要因についての分析が学術的にも実務的にも重要な研究テーマとなった。

GitHub には大量のプロジェクトが存在する。それらのデータは GitHub API により分析アプリケーションから直接取得できる。また、GitHub の活動データをアーカイブした学術的な Web サイト (BOA, GHTorrent など) からデータを取得することもできる。GitHub 出現後、複数のプロジェクトの比較や分類など、それらの方法で取得したデータ

を用いた定量的な分析がおこなわれるようになった。

本研究は、OSS 開発プロジェクトの成功を支援するために、貢献者獲得の要因を明確化することを目的とし、既存研究で可能性を示唆されていた代表的な要因である、インフルエンサー、プロジェクトの将来性、貢献ガイドラインについて、GitHub API などから取得したデータを用い定量的に評価するものである。

はじめに 2 章において、貢献者に関する先行研究に関して、「モチベーション」と「貢献者獲得」という観点から整理した。「モチベーション」については、多くの研究が E. L. デシ (Edward L. Deci) らによる「自己決定理論」をベースにしており、その用語をそのまま流用している。OSS に貢献する動機は、内発的動機、内在化された外発的動機、外発的動機の 3 つのカテゴリーに分類できる。それぞれについて多様な研究が存在するが、総括すると、実際にプロジェクトに貢献するという決断には、外発的動機、たとえば報酬とかキャリアなどが重要である。特に、単調な業務については外発的動機が重要であると考えられている。 しかしながら、創造性の高い作業では外発的動機は効果が薄く、また当初は外発的動機で参加したものの、継続して貢献していくうちに内発的動機、たとえば友人関係、趣味、娯楽などに内在化していくことも観察されている。また逆に、長期間貢献を継続してもらうためには、内発的動機が重要であると指摘されている。一方、「貢献者獲得」はライセンスタイプ、スポンサー、インフルエンサーなどが要因として研究されているが、まだ課題が多い。

3 章では、インフルエンサーについて分析をおこない、インフルエンサーの存在が貢献者獲得に有効であることを確認した。また、インフルエンサーの特定方法について複数のアルゴリズムを比較し、妥当性を検証した。具体的には、GitHub 上の仮想通貨プロジェクトからフォローネットワークを構築し、ネットワーク分析手法を用い分析をおこなった。

その結果、インフルエンサーの影響力と貢献者数との間に関連性が確認できた。これは、影響力のあるユーザーがプロジェクトへの貢献者を集めることに貢献していることを示唆している。また、3 つの中心性スコア（入次数、PageRank、HITS/Authority）を比較した結果、HITS/Authority スコアが、仮想通貨インフルエンサーの影響力の指標として最も適していることが確認できた。 HITS アルゴリズムは、仮想通貨フォローネットワークなどのような「ロックスター」（多くのフォロワーを集めるが、自分自身はほとんどフォローしない）ネットワークからドメイン固有のインフルエンサーを抽出するのに有

効であると考えられる。仮想通貨ドメインなどの専門的ドメインにおいて、インフルエンサーの存在が貢献者の獲得に有利に働くこと、影響力の指標として HITS アルゴリズムによる指標が有効であること、を実証できたことが本章での貢献である。

4 章では、プロジェクトの将来性について分析をおこない、プロジェクトの将来性が貢献者の獲得に有効であることを確認した。前章と同じく、分析データとしては GitHub 上の仮想通貨プロジェクトを使用した。仮想通貨は市場データが公開されており、時価総額の推移が取得できる。時価総額をプロジェクトの将来性を示す代理変数として、貢献者数との関連性を時系列分析手法で分析をおこなった。グレンジャー因果性検定の結果、1）時価総額から貢献者数への因果関係が存在する可能性があり、2）逆に貢献者数から時価総額への因果関係が存在しない、ことを確認できた。回帰分析の結果、時価総額が増加（減少）してから 2 か月後に貢献者の数が増加（減少）した。これは、プロジェクトの将来性（すなわち、時価総額）がプロジェクトへの参加を促すことを示唆していると考えられる。つまり時価総額の増加は、新しい貢献者の獲得につながると推測される。回帰分析の結果は、グレンジャー因果性検定の結果と一致しており、プロジェクトの将来性は貢献者の参加に影響を与えると結論づけた。

5 章では、GitHub の標準ファイルである貢献ガイドラインについて分析し、貢献ガイドラインの記載内容が貢献者数と関連性があるか検証した。また、貢献ガイドラインの記載内容を考察し、本来記載すべき内容を提言した。構造トピックモデルを用い記載内容をトピックに分解した後、各トピックと貢献者数との関連性を検証した。その結果、貢献者数と「不具合報告」トピックの出現確率との間には正の相関、「Git 更新操作」との間には負の相関が存在することがわかった。また、「環境構築」、「コーディング規約」、「Git 操作」、「不具合報告」トピックを貢献ガイドラインに記載することを提言した。ソフトウェア開発においてガイドラインは根幹をなす。ガイドラインが不備であると、ユーザーに不信感を与えるだけでなく、規則性が失われ不具合を生じやすくメンテナンスの負荷が増大する。本研究によりガイドラインの質が向上し、貢献者の獲得および生産性の向上に寄与することを期待している。

今後の研究課題については以下のとおりである。

1. 3 章、4 章で使用した仮想通貨のデータセットは金融情報を持つという特徴をもつ。また、仮想通貨プロジェクトの一覧が公開されており、根拠のないプロジェク

ト抽出という指摘 [CLC16] も回避できる。仮想通貨のデータセットは、今後さまざまな研究に利用できるだけでなく、OSS の貢献者に関する過去の研究の再検証にも適用可能である。すでに、このデータセットは認知され始めていて、利用した研究が始まっている [LAL+20]。

2. ５章では貢献ガイドラインを分析したが、本来貢献ガイドラインに書かれるべき内容を、他の標準ファイル（Readme や License ガイド) に記載しているプロジェクトも多く存在する。それらのファイルも分析対象として、記載内容が本来どのファイルに記述されるのが妥当か提言する。
3. ３章で影響力をフォロー数として分析したが、他のコラボレーション機能、「Fork」、「Watch」、「Star」などのデータも取得可能である。それらのデータも影響力の指標として活用できるか検証する。
4. 本研究では、インフルエンサー、将来性、貢献ガイドラインが貢献者獲得の要因となるか分析した。ただし、貢献者の種類まで踏み込んではいない。プロジェクトで必要としているのは、一回限りの貢献者ではなく、長期にわたって貢献してくれるユーザーである。長期貢献者への有効性についての分析は今後の課題である。

謝辞

修士論文の作成から３０年、全く畑違いの分野で博士課程に挑戦してようやくここまでたどり着くことができました。多くの方々にご指導、ご協力を賜りました。ここに感謝申し上げます。特に学術関係のことをすっかり忘れてしまった私に対し、一から懇切丁寧に指導して下さった吉田健一教授に深く感謝致します。 吉田先生に出会えたことは本当に幸運だったと思っています。また、副指導の倉橋節也教授、津田 和彦教授にも的確な助言を頂きました。両先生の講義で教えて頂いた分析方法が本論文の基礎になっています。本当にありがとうございました。最後に、長い間勝手な行動を許し温かく見守ってくれた、両親、家族に心より感謝します。

参考文献

[AA17] Ghadah Alamer and Sultan Alyahya. Open Source Software Hosting Platforms: A Collaborative Perspective's Review. *Journal of Software*, Vol. 12, No. 4, pp. 274–291, 2017.

[Abd05] Mohammad J Abdolmohammadi. Intellectual capital disclosure and market capitalization. *Journal of intellectual capital*, Vol. 6, No. 3, pp. 397–416, 2005.

[AGLM09] Dan Ariely, Uri Gneezy, George Loewenstein, and Nina Mazar. Large stakes and big mistakes. *The Review of Economic Studies*, Vol. 76, No. 2, pp. 451–469, 2009.

[AH02] Shaosong Ou Alexander Hars. Working for free? motivations for participating in open-source projects. *International journal of electronic commerce*, Vol. 6, No. 3, pp. 25–39, 2002.

[AHS14] Karan Aggarwal, Abram Hindle, and Eleni Stroulia. Co-evolution of project documentation and popularity within github. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, pp. 360–363, 2014.

[AL11] Oliver Alexy and Martin Leitner. A fistful of dollars: Are financial rewards a suitable management practice for distributed models of innovation? *European Management Review*, Vol. 8, No. 3, pp. 165–185, 2011.

[BA12] Jonathan M. Bischof and Edoardo M. Airoldi. Summarizing topical content with word frequency and exclusivity. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, Vol. 1,

pp. 201–208, 2012.

[Bar] Kevin R. Barnes. Contributing guidelines. `https://github.com/blog/1184-contributing-guidelines`.

[BHV16] Hudson Borges, Andre Hora, and Marco Tulio Valente. Understanding the factors that impact the popularity of github repositories. In *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 334–344. IEEE, 2016.

[BN12] Cinzia Battistella and Fabio Nonino. Open innovation web-based platforms: The impact of different forms of motivation on collaboration. *Innovation*, Vol. 14, No. 4, pp. 557–575, 2012.

[BNL20] Scott Brisson, Ehsan Noei, and Kelly Lyons. We Are Family: Analyzing Communication in GitHub Software Repositories and Their Forks. *SANER 2020 - Proceedings of the 2020 IEEE 27th International Conference on Software Analysis, Evolution, and Reengineering*, pp. 59–69, 2020.

[BNM+11] Christian Bird, Nachiappan Nagappan, Brendan Murphy, Harald Gall, and Premkumar Devanbu. Don't Touch My Code!: Examining the Effects of Ownership on Software Quality. *Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering*, p. 4, 2011.

[BS16] Ali Sajedi Badashian and Eleni Stroulia. Measuring user influence in GitHub. *Proceedings of the 3rd International Workshop on Crowd-Sourcing in Software Engineering - CSI-SE '16*, pp. 15–21, 2016.

[BSG+16] Kelly Blincoe, Jyoti Sheoran, Sean Goggins, Eva Petakovic, and Daniela Damian. Understanding the popular users: Following, affiliation influence and leadership on GitHub. *Information and Software Technology*, Vol. 70, pp. 30–39, 2016.

[CAH03] Kevin Crowston, Hala Annabi, and James Howison. Defining open source software project success. *Proceedings of the International Conference on Information Systems*, 06 2003.

[CF19] Todd A. Curry and Michael P. Fix. May it please the twitterverse: The use of Twitter by state high court judges. *Journal of Information Technology and Politics*, Vol. 16, No. 4, pp. 379–393, 2019.

[CHB+10] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, P Krishna Gummadi, et al. Measuring user influence in twitter: The million follower fallacy. *Icwsm*, Vol. 10, No. 10-17, p. 30, 2010.

[CLC16] Valerio Cosentino, Javier Luis, and Jordi Cabot. Findings from github: methods, datasets and limitations. In *Proceedings of the 13th International Conference on Mining Software Repositories*, pp. 137–141. ACM, 2016.

[CNF14] Christopher P. Cerasoli, Jessica M. Nicklin, and Michael T. Ford. Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, Vol. 140, No. 4, pp. 980–1008, 2014.

[CP16] Rudi Chen and Ivens Portugal. Analyzing factors impacting open-source project aliveness. *Univ. Waterloo, Waterloo, ON, Canada, Tech. Rep*, 2016.

[CSN09] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, Vol. 51, No. 4, pp. 661–703, 2009.

[CWHW12] Kevin Crowston, Kangning Wei, James Howison, and Andrea Wiggins. Free/Libre open-source software development. *ACM Computing Surveys*, Vol. 44, No. 2, pp. 1–35, 2012.

[DBR+18] Shishir Dubey, B Balaii, Dinesh Rao, Deepak Rao, et al. Data visualization on github repository parameters using elastic search and kibana. In *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 554–558. IEEE, 2018.

[DF95] Edward L Deci and Richard Flaste. *Why we do what we do: Understanding self-motivation.* Penguins Books, 1995.

[DG14] Elizabeth Dubois and Devin Gaffney. The Multiple Facets of Influ-

ence: Identifying Political Influentials and Opinion Leaders on Twitter. *American Behavioral Scientist*, Vol. 58, No. 10, pp. 1260–1277, 2014.

[DNRN13] Robert Dyer, Hoan Anh Nguyen, Hridesh Rajan, and Tien N Nguyen. Boa: A language and infrastructure for analyzing ultra-large-scale software repositories. In *2013 35th International Conference on Software Engineering (ICSE)*, pp. 422–431. IEEE, 2013.

[DR80] Edward L Deci and Richard M Ryan. Self-determination theory: When mind mediates behavior. *The Journal of mind and Behavior*, pp. 33–43, 1980.

[DSTH12] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. Social coding in github: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pp. 1277–1286. ACM, 2012.

[DWA03] Paul A David, Andrew Waterman, and Seema Arora. Floss-us the free/libre/open source software survey for 2003. *Stanford Institute for Economic Policy Research*, pp. 1–39, 2003.

[ESEZ19] Omar Elazhary, Margaret-Anne Storey, Neil Ernst, and Andy Zaidman. Do as i do, not as i say: Do contribution guidelines match the github contribution process? In *2019 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 286–290. IEEE, 2019.

[FG07] Chaim Fershtman and Neil Gandal. Open source software: Motivation and restrictive licensing. *International Economics and Economic Policy*, Vol. 4, No. 2, pp. 209–225, 2007.

[FN08] Yulin Fang and Derrick Neufeld. Understanding sustained participation in open source software projects. *Journal of Management Information Systems*, Vol. 25, No. 4, pp. 9–50, 2008.

[FPSF19] D. Fischer-Preßler, Carsten Schwemmer, and Kai Fischbach. Collective sense-making in times of crisis: Connecting terror management

theory with Twitter user reactions to the Berlin terrorist attack. *Computers in Human Behavior*, Vol. 100, pp. 138–151, 2019.

[Gho05] Rishab Aiyer Ghosh. Understanding free software developers: Findings from the floss study. *Perspectives on free and open source software*, Vol. 28, pp. 23–47, 2005.

[Git] GitHub. Celebrating nine years of github with an anniversary sale. `https://github.com/blog/2345-celebrating-nine-years-of-github-with-an-anniversary-sale`.

[GLM06] Rajdeep Grewal, Gary L. Lilien, and Girish Mallapragada. Location, location, location: How network embeddedness affects project success in open source systems. *Management Science*, Vol. 52, No. 7, pp. 1043–1056, 2006.

[GM11] Mathieu Goeminne and Tom Mens. Evidence for the Pareto principle in open source software activity. *CEUR Workshop Proceedings*, Vol. 708, pp. 74–82, 2011.

[GPD14] Georgios Gousios, Martin Pinzger, and Arie Van Deursen. An exploratory study of the pull-based software development model. *Proceedings - International Conference on Software Engineering*, No. 1, pp. 345–355, 2014.

[Gra69] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, Vol. 37, No. 3, p. 424, 1969.

[GS12] Georgios Gousios and Diomidis Spinellis. Ghtorrent: Github's data from a firehose. In *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*, pp. 12–21. IEEE, 2012.

[GS15] David Garcia and Frank Schweitzer. Social signals and algorithmic trading of bitcoin. *Royal Society open science*, Vol. 2, No. 9, p. 150288, 2015.

[GS17] Georgios Gousios and Diomidis Spinellis. Mining software engineering data from GitHub. In *Proceedings - 2017 IEEE/ACM 39th In-*

ternational Conference on Software Engineering Companion, ICSE-C 2017, pp. 501–502. Institute of Electrical and Electronics Engineers Inc., jun 2017.

[Ham94] James Douglas Hamilton. *Time series analysis*, Vol. 2. Princeton university press Princeton, NJ, 1994.

[Hen00] Beth A Hennessey. Rewards and creativity. In *Intrinsic and extrinsic motivation*, pp. 55–78. Elsevier, 2000.

[HJARTRBW94] F Hair Joseph, E Anderson Rolph, L Tatham Ronald, and C Black William. *Multivariate data analysis with readings*. Macmillan Publishing Company, 1994.

[HK03] Eric von Hippel and Georg von Krogh. Open source software and the "private-collective" innovation model: Issues for organization science. *Organization science*, Vol. 14, No. 2, pp. 209–223, 2003.

[HNH03] Guido Hertel, Sven Niedner, and Stefanie Herrmann. Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. *Research policy*, Vol. 32, No. 7, pp. 1159–1177, 2003.

[ICC15] Javier Cánovas Izquierdo, Valerio Cosentino, and Jordi Cabot. Attracting contributions to your github project. 2015.

[K+07] Greg KroahHartman, et al. Linux kernel development. In *Linux Symposium*, pp. 239–244. Citeseer, 2007.

[K+08] Max Kuhn, et al. Building predictive models in r using the caret package. *Journal of statistical software*, Vol. 28, No. 5, pp. 1–26, 2008.

[KGB+16] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M. German, and Daniela Damian. An in-depth study of the promises and perils of mining GitHub. *Empirical Software Engineering*, Vol. 21, No. 5, pp. 2035–2071, 2016.

[KHSW12] Georg Von Krogh, Stefan Haefliger, Sebastian Spaeth, and Martin W Wallin. Theory and Review Carrots and Rainbows : Motivation and

Social Practice in Open Source Software Development. Vol. 36, No. 2, pp. 649–676, 2012.

[KINY20] Naoki Kobayakawa, Mitsuyoshi Imamura, Kei Nakagawa, and Kenichi Yoshida. Impact of cryptocurrency market capitalization on open source software participation. *Journal of Information Processing*, Vol. 28, pp. 650–657, 2020.

[Kle99] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, Vol. 46, No. 5, pp. 604–632, 1999.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pp. 591–600. AcM, 2010.

[Kri02] Sandeep Krishnamurthy. Cave or community?: An empirical examination of 100 mature open source projects. *First Monday*, 2002.

[KY17] Naoki Kobayakawa and Kenichi Yoshida. How github contributing. md contributes to contributors. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 1, pp. 694–696. IEEE, 2017.

[KY19] Naoki Kobayakawa and Kenichi Yoshida. Study on influencers of cryptocurrency follow-network on github. In *Pacific Rim Knowledge Acquisition Workshop*, pp. 173–183. Springer, 2019.

[LAL+20] Lorenzo Lucchini, Laura Alessandretti, Bruno Lepri, Angela Gallo, and Andrea Baronchelli. From code to market: Network of developers and correlated returns of cryptocurrencies. 2020.

[LC03] Gwendolyn K Lee and Robert E Cole. From a firm-based to a community-based model of knowledge creation: The case of the linux kernel development. *Organization science*, Vol. 14, No. 6, pp. 633–649, 2003.

[LFCH13] MJ Lee, B Ferwerda, J Choi, and J Hahn. Github developers use rockstars to overcome overflow of news. *Chi Ea*, pp. 133–138, 2013.

[LRM14] Antonio Lima, Luca Rossi, and Mirco Musolesi. Coding together at scale: Github as a collaborative social network. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[LSR08] Christoph Lattemann, Stefan Stieglitz, and Wirtschafts-Sozialwissenschaftliche Reihe. Universität Potsdam Framework for Governance in Open Source Communities at Potsdam Framework for Governance in Open Source Communities. 2008.

[LT03] Josh Lerner and Jean Tirole. Some Simple Economics of Open Source. *The Journal of Industrial Economics*, Vol. 50, No. 2, pp. 197–234, 2003.

[LVH04] Karim R Lakhani and Eric Von Hippel. How open source software works: "free" user-to-user assistance. In *Produktentwicklung mit virtuellen Communities*, pp. 303–339. Springer, 2004.

[LW03] Karim R Lakhani and Robert G Wolf. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. 2003.

[MA00] M Lynne Markus and Brook Manville Carole E Agres. What makes a virtual organization work? *MIT Sloan Management Review*, Vol. 42, No. 1, p. 13, 2000.

[MB12] Teng Sheng Moh and Surya Bhagvat. Clustering of technology tweets and the impact of stop words on clusters. *Proceedings of the Annual Southeast Conference*, pp. 226–231, 2012.

[MF70] Burton G Malkiel and Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, Vol. 25, No. 2, pp. 383–417, 1970.

[MWT+11] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, No. 2, pp. 262–272, 2011.

[Nak08] Satoshi Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash System. *Www.Bitcoin.Org*, p. 9, 2008.

[New05] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, Vol. 46, No. 5, pp. 323–351, 2005.

[NYN+02] Kumiyo Nakakoji, Yasuhiro Yamamoto, Yoshiyuki Nishinaka, Kouichi Kishida, and Yunwen Ye. Evolution patterns of open-source software systems and communities. *2nd International Workshop on Principles of Software Evolution (IWPSE 2002)*, Vol. 2002, No. January, p. 76, 2002.

[OF00] Margit Osterloh and Bruno S Frey. Motivation, knowledge transfer, and organizational forms. *Organization science*, Vol. 11, No. 5, pp. 538–550, 2000.

[ON08] Shaul Oreg and Oded Nov. Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values. *Tayler Computers in Human Behavior*, Vol. 24, No. 5, pp. 2055–2073, 2008.

[OO07] Chitu Okoli and Wonseok Oh. Investigating recognition-based performance in an open content community: A social capital perspective. *Information and Management*, Vol. 44, No. 3, pp. 240–252, 2007.

[PG17] R. C. Phillips and D. Gorse. Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7, Nov 2017.

[PSG16] Gustavo Pinto, Igor Steinmacher, and Marco Aurélio Gerosa. More Common Than You Think: An In-depth Study of Casual Contributors. *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, No. 1, pp. 112–123, 2016.

[PSL+13] Raphael Pham, Leif Singer, Olga Liskin, Fernando Figueira Filho, and Kurt Schneider. Creating a shared understanding of testing culture on a social coding site. In *2013 35th International Conference on*

Software Engineering (ICSE), pp. 112–121. IEEE, 2013.

[PTT+19] Gede Artha Azriadi Prana, Christoph Treude, Ferdian Thung, Thushari Atapattu, and David Lo. Categorizing the Content of GitHub README Files. *Empirical Software Engineering*, Vol. 24, No. 3, pp. 1296–1327, 2019.

[Ray01] Eric S Raymond. The cathedral and the bazaar: Musings on linux and open source by an accidental revolutionary. sebastopol, ca: Oreilly media, 2001.

[RD00] Richard M Ryan and Edward L Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, Vol. 55, No. 1, p. 68, 2000.

[RHS06] Jeff Roberts, Il-Horn Hann, and Sandra A Slaughter. Understanding the Motivations, Participation, and Performance of Open Source Software Developers: A ... *Management Science*, 2006.

[RSA16] Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, Vol. 111, No. 515, pp. 988–1003, 2016.

[RST19] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. Stm: An R package for structural topic models. *Journal of Statistical Software*, Vol. 91, No. 2, 2019.

[RSTA13] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, and Eduardo M. Airoldi. The structural topic model and applied social science. *NIPS 2013 Workshop on Topic Models*, pp. 2–5, 2013.

[SAM06] Katherine J. Stewart, Anthony P. Ammeter, and Likoebe M. Maruping. Impacts of license choice and organizational sponsorship on user interest and development activity in open source software projects. *Information Systems Research*, Vol. 17, No. 2, pp. 126–144, 2006.

[Sha06] Sonali K. Shah. Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development. *Management Science*,

Vol. 52, No. 7, pp. 1000–1014, 2006.

[Sta06] Richard Stallma. The Free Software Movement and the GNU/Linux Operating System. pp. 426–426, 2006.

[TBLJ13] Ferdian Thung, Tegawende F. Bissyandé, David Lo, and Lingxiao Jiang. Network structure of social coding in GitHub. *Proceedings of the European Conference on Software Maintenance and Reengineering, CSMR*, pp. 323–326, 2013.

[TDH13] Jason Tsay, Laura Dabbish, and James D. Herbsleb. Social media in transparent work environments. *2013 6th International Workshop on Cooperative and Human Aspects of Software Engineering, CHASE 2013 - Proceedings*, pp. 65–72, 2013.

[TDH14] Jason Tsay, Laura Dabbish, and James Herbsleb. Influence of social and technical factors for evaluating contribution in GitHub. *36th International Conference on Software Engineering*, pp. 356–366, 2014.

[VH01] Eric Von Hippel. Learning from open-source software. *MIT Sloan management review*, Vol. 42, No. 4, pp. 82–86, 2001.

[VPR+15] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark G.J. van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. Gender and Tenure Diversity in GitHub Teams. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pp. 3789–3798, 2015.

[WMWS19] Jing Wang, Xiangxin Meng, Huimin Wang, and Hailong Sun. An online developer profiling tool based on analysis of gitlab repositories. In *CCF Conference on Computer Supported Cooperative Work and Social Computing*, pp. 408–417. Springer, 2019.

[YK03] Yunwen Ye and Kouichi Kishida. Toward an understanding of the motivation of open source software developers. In *25th International Conference on Software Engineering, 2003. Proceedings.*, pp. 419–429. IEEE, 2003.

[YKM+16] Kazuhiro Yamashita, Yasutaka Kamei, Shane McIntosh, Ahmed E.

Hassan, and Naoyasu Ubayashi. Magnet or Sticky? Measuring Project Characteristics from the Perspective of Developer Attraction and Retention. *Journal of Information Processing*, Vol. 24, No. 2, pp. 339–348, 2016.

[YMKU14] Kazuhiro Yamashita, Shane McIntosh, Yasutaka Kamei, and Naoyasu Ubayashi. Magnet or sticky? an OSS project-by-project typology. *Proceedings of the 11th Working Conference on Mining Software Repositories - MSR 2014*, pp. 344–347, 2014.

[YYWW14] Yue Yu, Gang Yin, Huaimin Wang, and Tao Wang. Exploring the Patterns of Social Behavior in GitHub. *Proceedings of the 1st International Workshop on Crowd-based Software Development Methods and Technologies*, pp. 31–36, 2014.

[ZS10] Wenbin Zhang and Steven Skiena. Trading strategies to exploit blog and news sentiment. *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Vol. d, pp. 375–378, 2010.

[中谷 07] 中谷素之. 学ぶ意欲を育てる人間関係づくり: 動機づけの教育心理学, 2007.

関連業績リスト

第 3 章 Naoki Kobayakawa and Kenichi Yoshida, "Study on Influencers of Cryptocurrency Follow-Network on GitHub", Pacific Rim Knowledge Acquisition Workshop. Lecture Notes in Artificial Intelligence, vol 11669. Springer, Cham, 2019. p. 173-183.

第 4 章 Naoki Kobayakawa, Mitsuyoshi Imamura, Kei Nakagawa and Kenichi Yoshida, "Impact of Cryptocurrency Market Capitalization on the Open Source Software Participation",Journal of Information Processing, 2020, Volume 28, Pages 650-657

第 5 章 Naoki Kobayakawa and Kenichi Yoshida, "How GitHub Contributing.md Contributes to Contributors",2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). IEEE, 2017. p. 694-696.