CRANパッケージ名とGitHub上Rソフトウェアの名前重複の実態分析

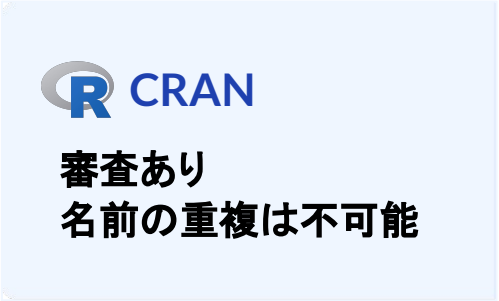
信州大学 工学部 電子情報システム工学科 4年
石田晃聖

Comprehensive R Archive Network (CRAN)

Rの本体や様々なRパッケージの保管、公開を行うサーバーのネットワーク。

特徴：

- パッケージはCRAN運営の審査を経て公開される。
- パッケージ名の重複は不可能。



背景と目的

GitHubにはCRANパッケージと同じ名前を持つリポジトリが存在している。

同名リポジトリが存在する要因として、人気パッケージの他の開発者による模倣や、GitHubに既にある名前をCRANパッケージ名に用いるなどの要因が考えられる。

研究目的

GitHubにおける同名リポジトリの実態とCRANパッケージのダウンロード数との関連を明らかにする。

データ収集

CRANデータ

- **対象**: CRAN全パッケージ
- **収集手法**: CRANDB API, CRANLogs API
- **取得情報**: パッケージ名, 作成日, 説明文, URL, 月間ダウンロード数
- **総数**: 22,781件

GitHubデータ

- **対象**: CRANパッケージと同名のRリポジトリ
- **収集手法**: GitHub REST API, GraphQL API
- **取得情報**: リポジトリ名, 作成日, 説明文, URL
- **総数**: 32,884件

Research Questions

RQ1. GitHub上の同名リポジトリはどの程度存在しているか

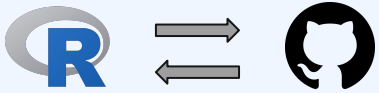
RQ2. 同名リポジトリの有無とCRANダウンロード数はどのように関連しているか

RQ3. 同名リポジトリとCRANダウンロード数の時間的関係はどうなっているか

RQ1　CRANパッケージの分類基準

CRAN公式との関連度：

公式へ誘導あり



CRANから公式リポジトリへのURLがある。
or 同名GitHubリポジトリにCRANへのURLがある。

公式へ誘導なし

CRANから公式リポジトリへのURLがない。
& 同名GitHubリポジトリにCRANへのURLがない。

同名リポジトリ（誘導あり除く）の作成時期：

既に同名あり
CRAN公開前に既に同名リポジトリが作成されている。

後から同名あり
CRAN公開後に初めて同名リポジトリが作成される。

同名なし
同名リポジトリが存在しない。

RQ1　結果

- CRANパッケージの57.5%がCRAN公式への誘導があるリポジトリを持つ。
- 「既に同名あり」「後から同名あり」がそれぞれ約4000件存在する。

表1 同名リポジトリ出現時期別 CRANパッケージ数

分類	既に同名あり	後から同名あり	同名なし	合計パッケージ数
公式へ誘導あり	1,990	1,678	9,424	13,092
公式へ誘導なし	2,023	2,288	5,378	9,689

RQ2　結果

公式へ誘導があるケース、後から同名GitHubリポジトリが作成されるケースはCRANパッケージのダウンロード数が多い傾向にある。

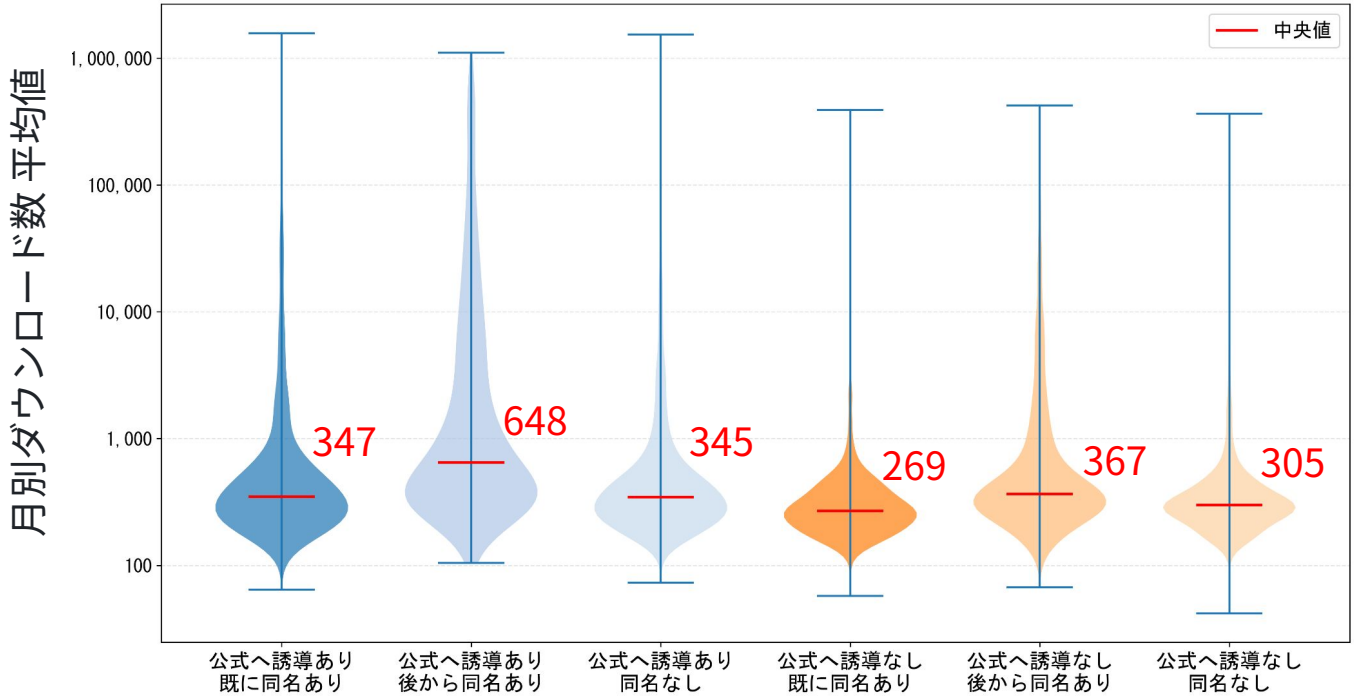


図1 6グループ毎のダウンロード数の分布

RQ3　結果

「後から同名あり」パッケージは、公開後約20ヶ月以降からダウンロード数が他2グループを超える傾向にある。

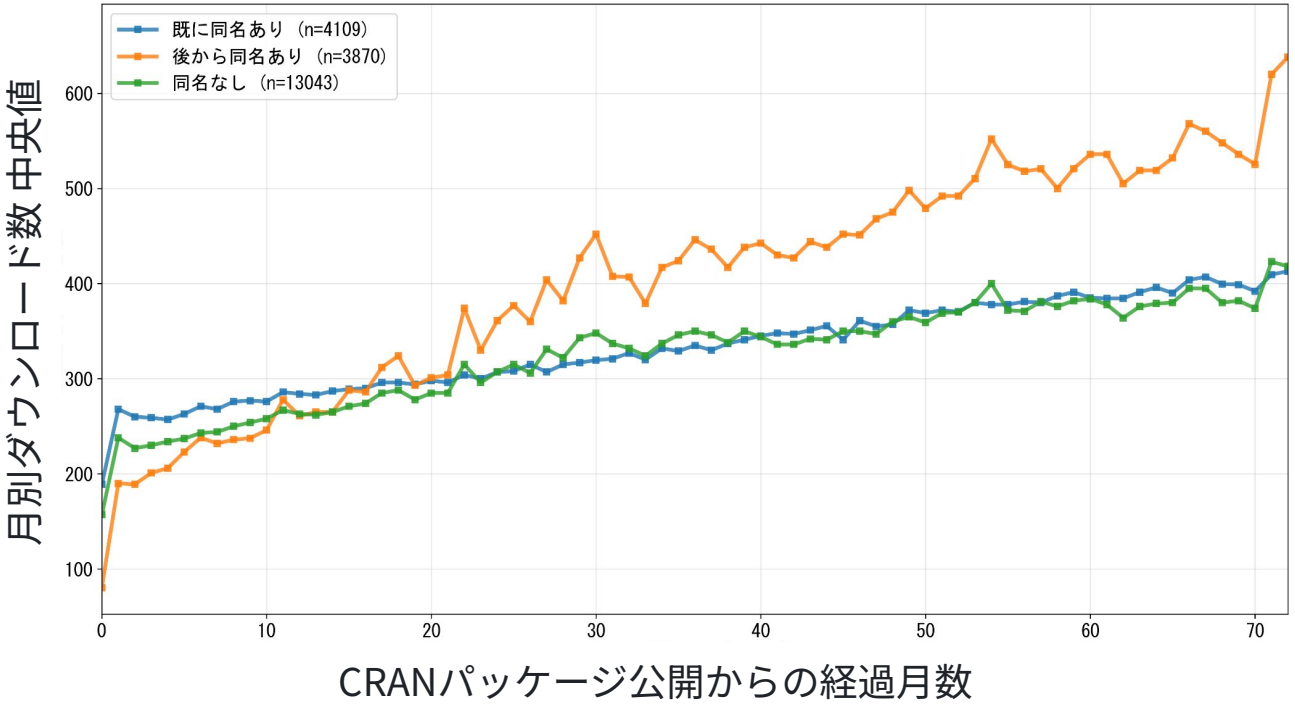


図2 CRANパッケージ公開後 時系列ダウンロード数(中央値)

RQ3　結果

同名GitHubリポジトリ作成後、CRANダウンロード数は上昇する傾向にある。

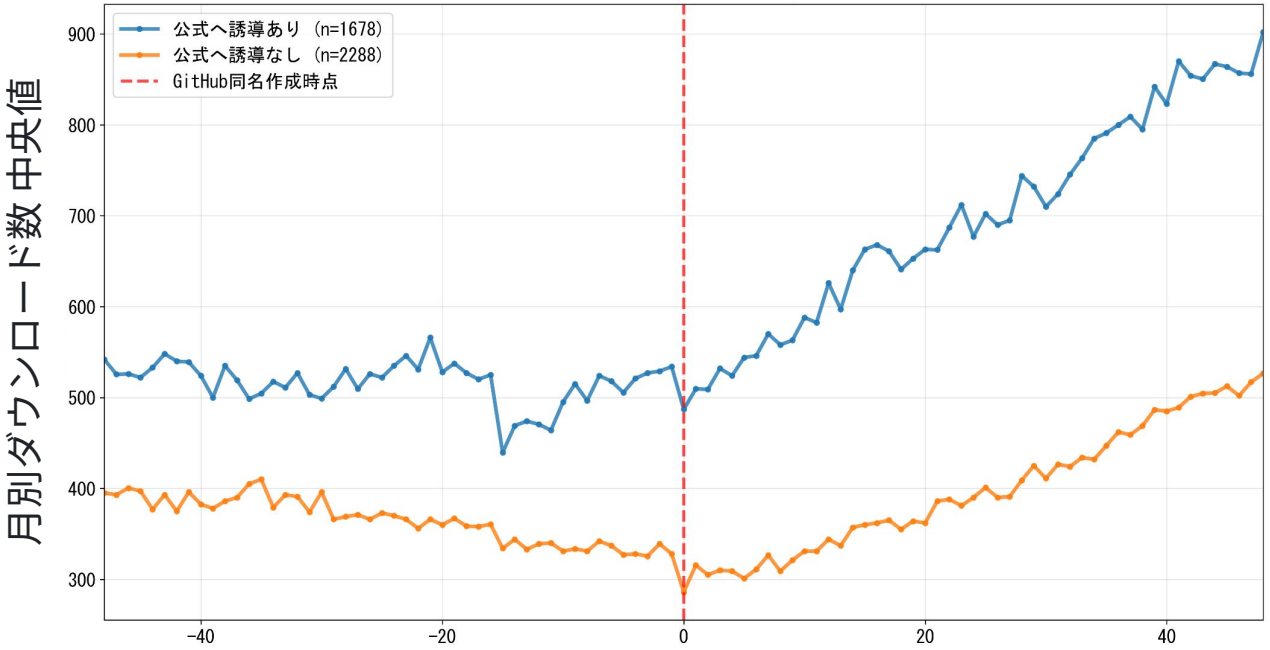


図3 「後から同名あり」時系列ダウンロード数(中央値)

分析結果まとめ

RQ1:

- 35%のパッケージがCRAN公式へ誘導が無い同名GitHubリポジトリを持つ。

RQ2:

- CRANへ誘導があるGitHubリポジトリが存在するとダウンロード数が増加する。
- 「後から同名あり」のケースは他と比較してCRANパッケージのダウンロード数が多い。

RQ3:

- 「後から同名あり」のケースでは、最初の同名GitHubリポジトリの作成後、CRANパッケージのダウンロード数が増加する傾向にある。

考察

- 公式のリポジトリや、CRANURLを記載したGitHubリポジトリがあるとダウンロード数が増加する。

  →GitHubからCRANパッケージへの誘導があると、ユーザーはCRANのサイトを閲覧しインストールを行う。

- 後発で同名GitHubリポジトリが作成されるとCRANパッケージのダウンロード数が増加する。

  →同名GitHubリポジトリの存在がユーザーにリポジトリ名の検索を促し、CRANサイトやCRANのURLを記載した他リポジトリの発見に繋がる。

まとめ

- CRANと関わりが無いように見える同名リポジトリが35%のパッケージで存在する。

- CRANパッケージへの誘導がある場合や、CRAN公開後に後発で同名GitHubリポジトリが作成された場合、CRANパッケージのダウンロード数が増加する現象が確認できた。

- 統計的手法を用いた因果関係の分析や、パッケージやリポジトリ単体のファイル構造・スクリプト内容の類似度による傾向の変化の分析を行いたい。