The Signals that Potential Contributors Look for When Choosing Open-source Projects
Goal: Increase participation in Open Source

• To sustain open source projects that need constant supply of effort

• To provide opportunities to learn / develop skills / showcase talent

• To increase diversity

• …
Participation in Open Source has a long history of scholarship

1. discover a relevant project
2. find an opportunity to contribute
3. make a first contribution
4. demonstrate commitment over time
5. be recognized as core contributor
Yet, little is known about how people choose which project to contribute to

Especially for people who contribute for fun, to learn, etc.

discover a relevant project
find an opportunity to contribute
make a first contribution
demonstrate commitment over time
be recognized as core contributor
Project discovery is especially relevant on transparent platforms like GitHub.
GitHub users rely on signals for many inferences
(Dabbish et al., 2012)
GitHub users rely on signals for many inferences

(Trockman et al., 2018)

Status

- Tests passing
- npm v4.3.1
- gem v4.3.1
- meteor twbs:bootstrap
- packagist v4.3.1
- nuget v4.3.1
- peer dependencies up to date
- dev dependencies up to date
- Coverage 94%
- CSS gzip size 21.2 KB
- JS gzip size 17.6 KB
- BrowserStack completed
- backers 27
- sponsors 18

What's included

Within the download you'll find the following directories and files, logically grouping common assets and providing both compiled and minified variations. You'll see something like this:
How do people use signals when choosing a GitHub project to contribute to?
Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>README</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contributing guideline</td>
</tr>
<tr>
<td>Community</td>
<td>Impolite language</td>
</tr>
<tr>
<td></td>
<td>Issue template</td>
</tr>
<tr>
<td>Popularity</td>
<td>Stars</td>
</tr>
<tr>
<td></td>
<td>Recent commits</td>
</tr>
</tbody>
</table>
Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>README</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contributing guideline</td>
</tr>
<tr>
<td>Community</td>
<td>Impolite language</td>
</tr>
<tr>
<td></td>
<td>Issue template</td>
</tr>
<tr>
<td>Popularity</td>
<td>Stars</td>
</tr>
<tr>
<td></td>
<td>Recent commits</td>
</tr>
</tbody>
</table>
Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>README</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contributing guideline</td>
</tr>
<tr>
<td>Community</td>
<td>Impolite language</td>
</tr>
<tr>
<td></td>
<td>Issue template</td>
</tr>
<tr>
<td>Popularity</td>
<td>Stars</td>
</tr>
<tr>
<td></td>
<td>Recent commits</td>
</tr>
</tbody>
</table>

**Contributing**

Please read through our contributing guidelines. Included are development.

Moreover, if your pull request contains JavaScript patches or features, they should conform to the Code Guide, maintained by Microsoft.

Editor preferences are available in the editor config for easy use. Check it out at https://editorconfig.org/.
### Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>README</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contributing guideline</td>
</tr>
<tr>
<td>Community</td>
<td>Impolite language</td>
</tr>
<tr>
<td></td>
<td>Issue template</td>
</tr>
<tr>
<td>Popularity</td>
<td>Stars</td>
</tr>
<tr>
<td></td>
<td>Recent commits</td>
</tr>
</tbody>
</table>

![Image of repository with files listed](image-url)

- `build`: YES! F**K YOU WEBPACK
- `dist`: Precompile assets
- `docs`: Experimenting with component data tables
- `public`: Rederubfne
- `src`: Move redirects again
Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>README</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contributing guideline</td>
</tr>
<tr>
<td>Community</td>
<td>Impolite language</td>
</tr>
<tr>
<td></td>
<td>Issue template</td>
</tr>
<tr>
<td>Popularity</td>
<td>Stars</td>
</tr>
<tr>
<td></td>
<td>Recent commits</td>
</tr>
</tbody>
</table>

```yaml
## I'm submitting a...

Please search GitHub for a similar issue or PR before submitting

<pre><code>
[ ] Regression (a behavior that used to work and stopped working in a new release)
[ ] Bug report
[ ] Feature request
[ ] Documentation issue or request
</code></pre>
Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>README</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contributing guideline</td>
</tr>
<tr>
<td>Community</td>
<td>Impolite language</td>
</tr>
<tr>
<td></td>
<td>Issue template</td>
</tr>
<tr>
<td>Popularity</td>
<td>Stars</td>
</tr>
<tr>
<td></td>
<td>Recent commits</td>
</tr>
</tbody>
</table>
Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>README</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contributing guideline</td>
</tr>
<tr>
<td>Community</td>
<td>Impolite language</td>
</tr>
<tr>
<td></td>
<td>Issue template</td>
</tr>
<tr>
<td>Popularity</td>
<td>Stars</td>
</tr>
<tr>
<td></td>
<td>Recent commits</td>
</tr>
</tbody>
</table>

Contributions to master, excluding merge commits

![Graph showing contributions to master, excluding merge commits]
Methodology
A mixed-method study design - Qualitative

Interviews with 15 developers

Qualitative analysis
Qualitative: Review and recommend projects

Pre-selected projects
Shown a set of NPM projects’ GitHub pages

Think-aloud interviews
Describe how they evaluate each project

Persona
Decide for each project, if they would recommend it to Jane

Jane,
Bachelor degree in Computer Science
First-time contributor to npm
Interviewees came from diverse backgrounds

Full-time developers: 33% YES, 67% NO

Gender: 40% FEMALE, 60% MALE

Years of OSS experience:
- < 1 year: 7%
- 1-5 years: 13%
- 6-10 years: 27%
- 10+ years: 53%
A mixed-method study design - Quantitative

- Interviews with 15 developers
- Qualitative analysis
- Quantitative hypotheses
- Test signal attractiveness
Quantitative: GitHub data mining

Mining trace data

~10,000 npm packages from GitHub and GHTorrent

Operationalization

14 quantitative measurements

Logistic regression

Associations between signals and newcomers

Jan 2018, Jun 2018, Sep 2018
## Result highlights:

Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>Interviews</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>README</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Contributing guideline</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Community</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impolite language</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Issue template</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Popularity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stars</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Recent commits</td>
<td>⬤</td>
<td>⬤</td>
</tr>
</tbody>
</table>
Three main classes of signals

Quality
Community
Popularity
Interviews:

“README is a project’s welcome mat”
“the README is most important. It should describe **without** having to navigate away from that page the key information people need"
A nice README should have:

**a table of contents**

Table of contents

- Quick start
- Status
- What’s included
- Bugs and feature requests
- Documentation
- Contributing
- Community
A nice README should have:

descriptions and instructions
A nice README should have:

contact information
A nice README should have:

**code quality badges**
Logistic regression:
Counting the number of headers in README
Logistic regression:
A well-structured README correlates to higher likelihood of attracting newcomers
Interviews:

Contributing file “suggests they have some experience with handling new contributors”
A nice contributing guideline should be: prominent
A nice contributing guideline should:

explain the GitHub jargon

Using the issue tracker

The issue tracker is the preferred channel for bug reports, features requests and submitting pull requests, but please respect the following restrictions:

- Please do not derail or troll issues. Keep the discussion on topic and respect the opinions of others.
A nice contributing guideline should: mention ways to contribute other than code

Using the issue tracker

The issue tracker is the preferred channel for bug reports, features requests and submitting pull requests, but please respect the following restrictions:

- Please do not derail or troll issues. Keep the discussion on topic and respect the opinions of others.
Logistic regression:
The presence of contributing guidelines correlates with lower likelihood of attracting newcomers.
Logistic regression:

Contributing Guidelines and amount of project activity interact: help vs overhead
Three main classes of signals

- Quality
- Community
- Popularity
Interviews:
The use of language signals the community’s openness

Hi @g0rm4l 0x, could you share a minimal code snippet with the text input that produces this error, so we can more easily diagnose it? Thanks in advance!

Contributing

1. Search the issues, if it is not already there, add one.
2. Fork the repository
3. Code Code Code
4. Submit a pull request
5. Wait for some nice guy to review and merge
Logistic regression:
Measuring impoliteness in conversations

lol dude, that's not how this works.

@complin and I have clearly explained in #1809 why we won't add this to Compose. Opening a new issue won't change that.
Logistic regression:
Impolite language correlates with lower likelihood of attracting new contributors

Only for experienced contributors!

First-time GitHub contributors:
Interviews:

Having a template signals that “there’s a good structure for contributing to [this project]”

---

```markdown
## I'm submitting a...

Please search GitHub for a similar issue or PR before submitting

<pre><code>
[ ] Regression (a behavior that used to work and stopped working in a new release)
[ ] Bug report
[ ] Feature request
[ ] Documentation issue or request

(http://github.com/twbs/bootstrap/blob/master/)

**Bug reports must include:**

- Operating system and version (Windows, macOS,...
- Browser and version (Chrome, Firefox, Safari, IE)
- [Reduced test case](http://css-tricks.com/reduced)
- [CodePen](http://codepen.io/) or [JS Bin](http://jsbin.com/)
```
Logistic regression:
Having a template correlates with higher likelihood of attracting new contributors

!Only for experienced contributors!

First-time GitHub contributors:
Three main classes of signals

Quality

Community

Popularity
Interviews:
The number of stars signals popularity
Interviews: Three different opinions

“Everyone uses [project X]. If you contribute to it your change is gonna have a huge impact.”

“If it has tons of stars, and there weren’t that many contributors, I would think they weren’t friendly to new people”.

“I wouldn’t worry about popularity, because a less popular project gives you more self-efficacy that forces you really to look at things, ... try everything out.”
Logistic regression:
The number of stars correlates strongly with higher likelihood of attracting new contributors
Interviews:
Recency of commits signals that the repo is active
Logistic regression:
More recent commits correlates with higher likelihood of attracting new contributors

!Only for experienced contributors!

First-time GitHub contributors:
## Result highlights:

Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>Interviews</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>README</td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>Contributing guideline</td>
<td>🟡</td>
<td>🟢</td>
</tr>
<tr>
<td>Community</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impolite language</td>
<td>🟡</td>
<td>🟡</td>
</tr>
<tr>
<td>Issue template</td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>Popularity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stars</td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>Recent commits</td>
<td>🟢</td>
<td>🟢</td>
</tr>
</tbody>
</table>
What can we do?
New signals (maybe in the form of badges)

Logistic regression:
More recent commits correlates with **higher** likelihood of attracting new contributors

Logistic regression:
Impolite language correlates with **lower** likelihood of attracting new contributors

**maintained** no! (as of 2018)
commit activity 457/month
issue resolution 3 h
updated today

**commitizen** friendly
PRs welcome
help-wanted
!Only for experienced contributors!
“Everyone uses [project X]. If you contribute to it your change is gonna have a huge impact.”

“a less popular project gives you more self-efficacy that forces you really to look at things, ... try everything out.”
Many signals used when choosing projects

Yet, little is known about how people choose which project to contribute to
Especially for people who contribute for fun, to learn, etc.

A mixed-method study design

- Interviews with 15 developers
- Qualitative analysis
- Quantitative hypotheses
- Test signal attractiveness

Result highlights:
Many signals used when choosing projects

<table>
<thead>
<tr>
<th>Quality</th>
<th>Interviews</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>README</td>
<td>🍀</td>
<td>🍀</td>
</tr>
<tr>
<td>Contributing guideline</td>
<td>🍀</td>
<td>🍀</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Community</th>
<th>Interviews</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impolite language</td>
<td>🍀</td>
<td>🍀</td>
</tr>
<tr>
<td>Issue template</td>
<td>🍀</td>
<td>🍀</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Popularity</th>
<th>Interviews</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stars</td>
<td>🍀</td>
<td>🍀</td>
</tr>
<tr>
<td>Recent commits</td>
<td>🍀</td>
<td>🍀</td>
</tr>
</tbody>
</table>

New signals (maybe in the form of badges)

- Maintained (no! (as of 2018))
- Commit activity 457/month
- Issue resolution 3 h
- Updated today
- PRs welcome
- Help-wanted

Acknowledgements:

Contact: Sophie Qiu
@sophiehsqq
hsqq@cmu.edu

@asarma
Anita.Sarma@oregonstate.edu

@b_vasilescu
vasilescu@cmu.edu

Code and data:
DOI 10.5281/zenodo.3371186