

10th ANNIVERSARY



Sustaining Open Source Digital Infrastructure

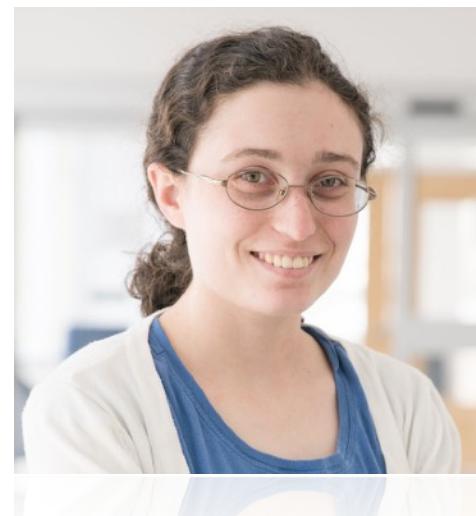
Bogdan Vasilescu
@b_vasilescu



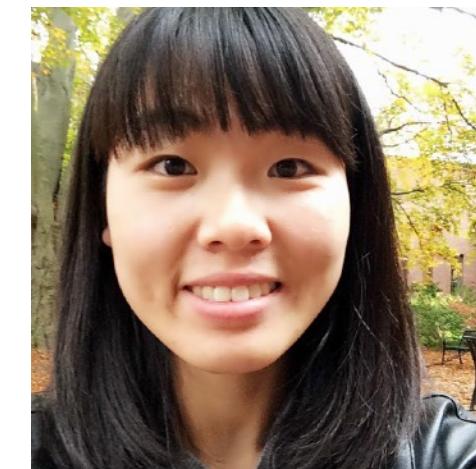
Acknowledgements



Courtney Miller



Anita Brown



Michelle Cao



Jim Herbsleb



Christian Kästner



David Widder



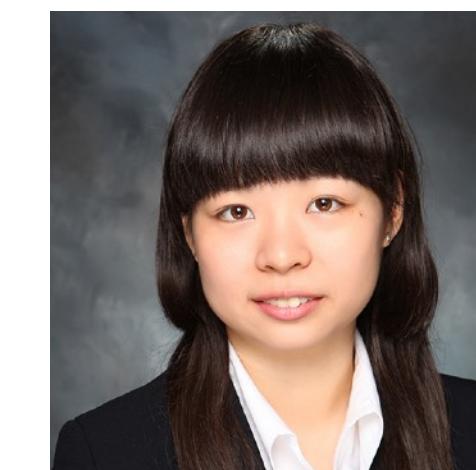
Anita Sarma



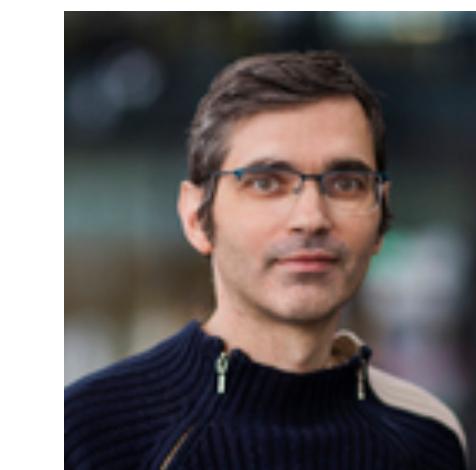
Audris Mockus



Alex Nolte



Sophie Qiu



Alex Serebrenik



Marat Valiev



Laura Dabbish



Lily Li



National Science
Foundation



Alfred P. Sloan
FOUNDATION

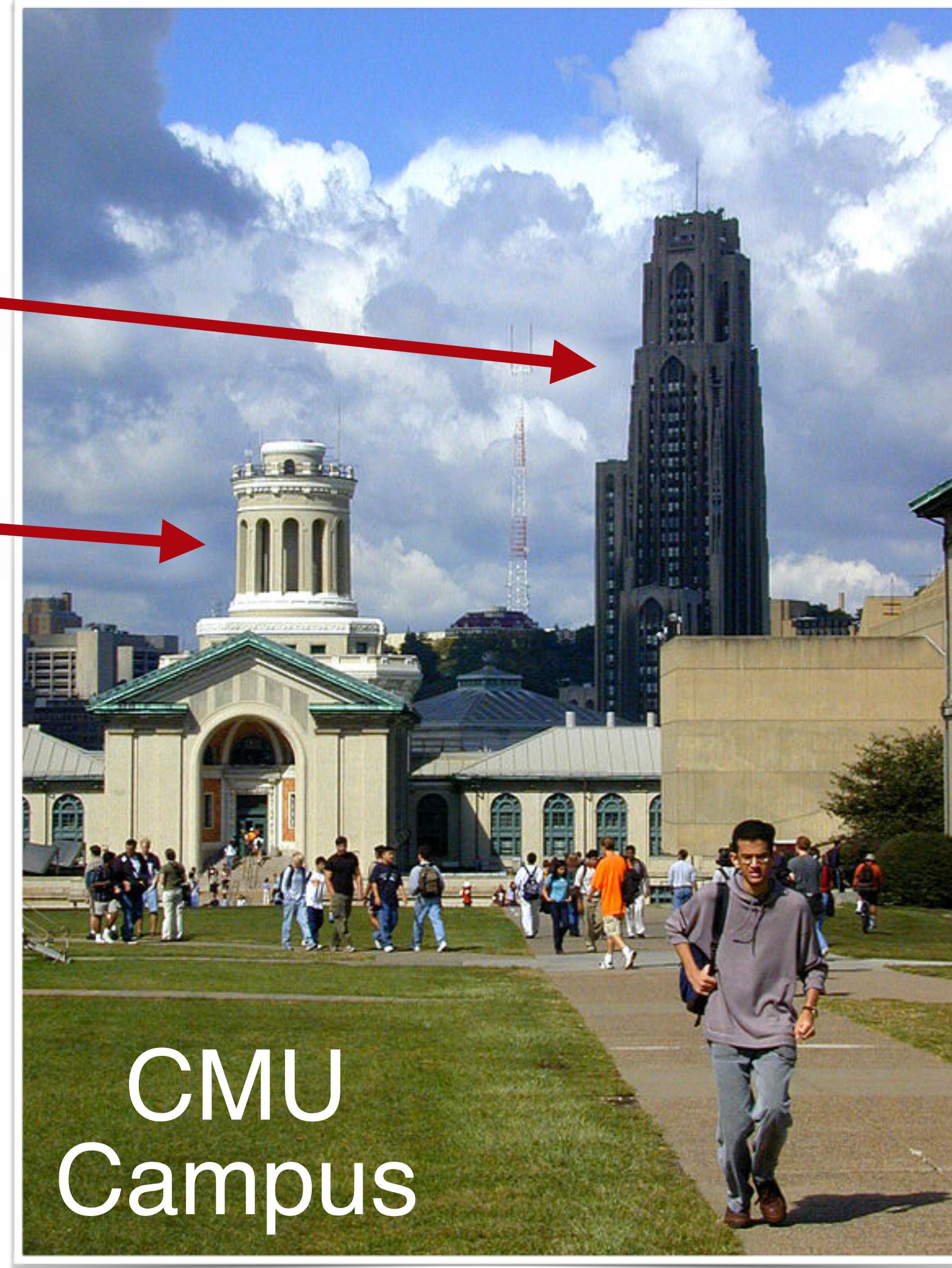


FORDFOUNDATION

I'm
here
to
learn
from
you

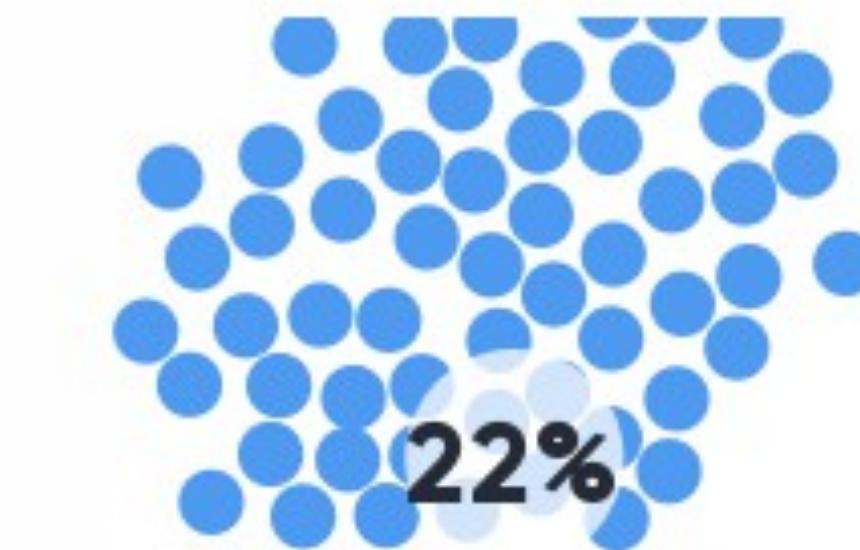
Ivory
tower #2

Ivory
tower #1

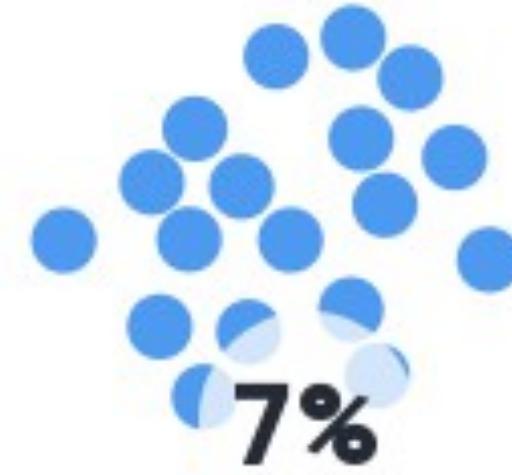


CC-BY-SA-2.0 https://commons.wikimedia.org/wiki/File:CMU_campus_Cathedral_Learning_background.jpg

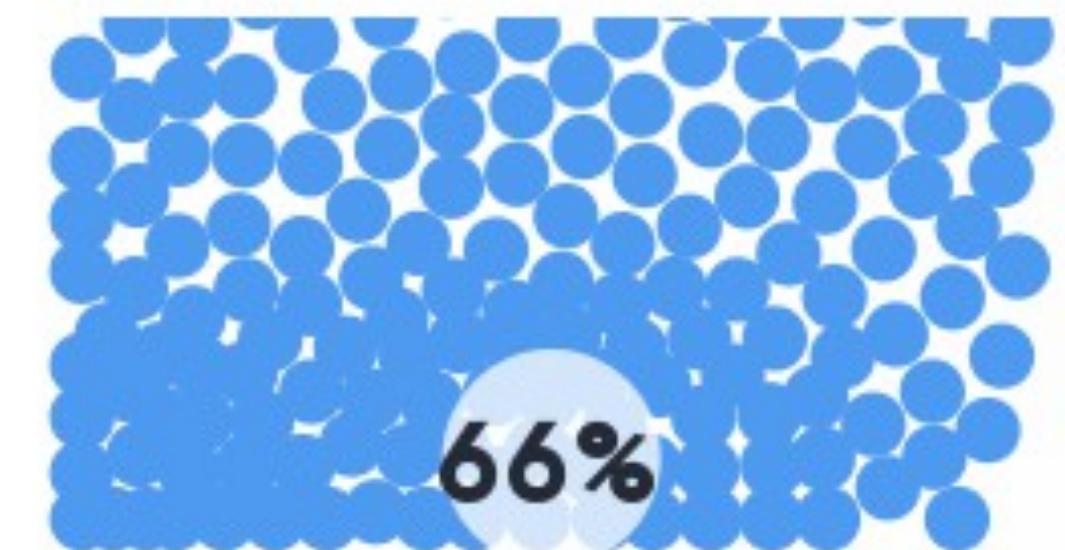
How do you engage with open source?



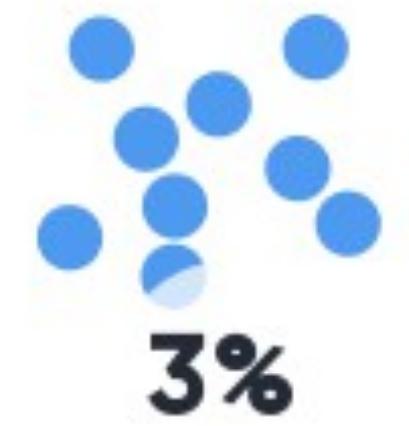
I maintain one or more open source projects



I regularly contribute to open source projects



I regularly use open source projects in proprietary work



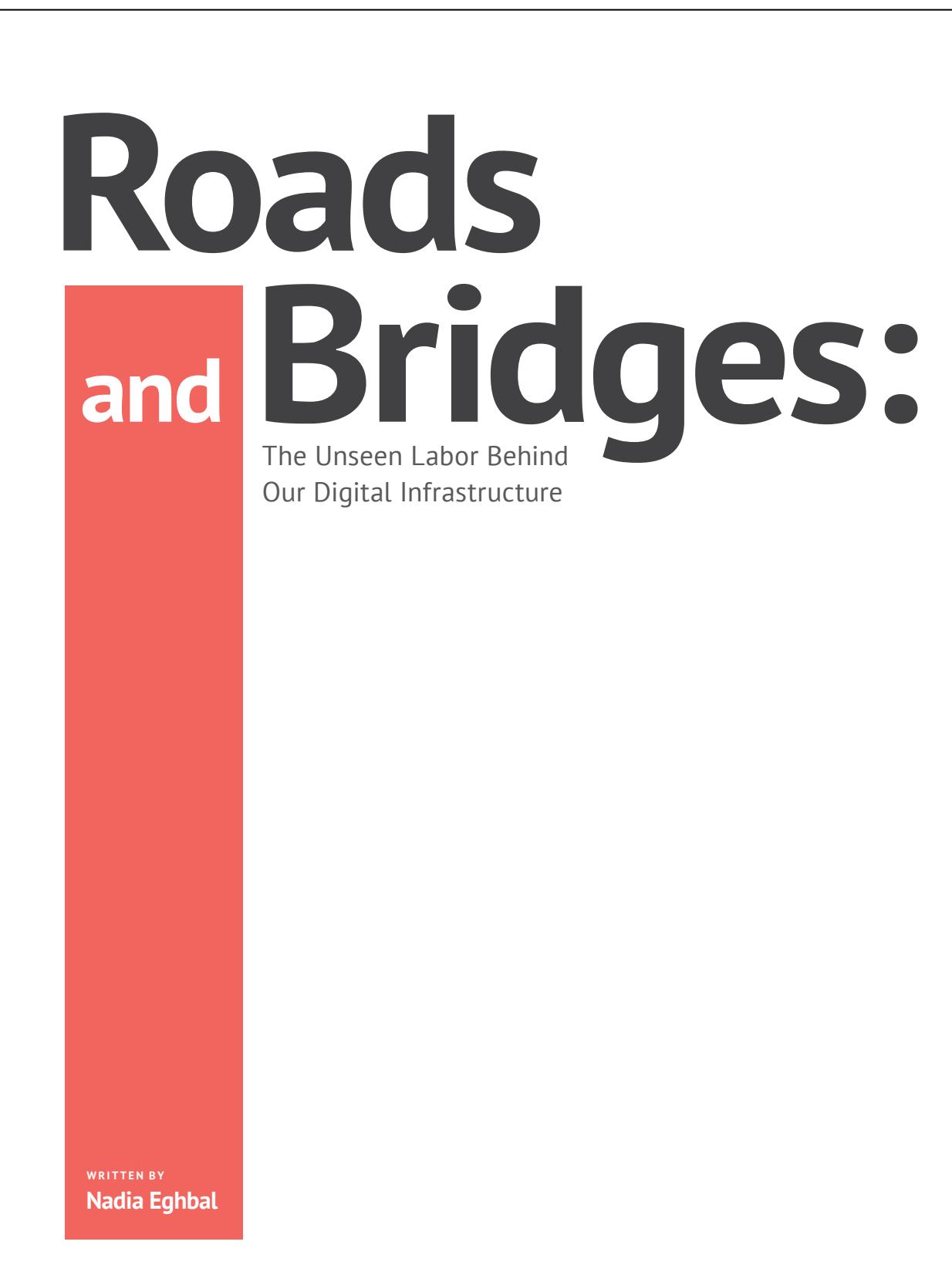
I rely primarily on closed source



I don't code

273

Open Source as digital infrastructure: Needs regular upkeep and maintenance



- Everybody uses open source code:
 - Fortune 500 companies
 - major software companies
 - startups
 - government
 - ...
- If undermaintained:
 - Risks for downstream users
 - Slows down innovation
 - ...

NPM ERR!

How one programmer broke the internet by deleting a tiny piece of code

By Keith Collins • March 27, 2016

A screenshot of a terminal window showing a bug in the leftpad.js file. The code is as follows:

```
1 module.exports = leftpad;
2 function leftpad (str, len, ch) {
3     str = String(str);
4     var i = -1;
5     if (!ch && ch !== 0) ch = ' ';
6     len = len - str.length;
7     while (++i < len) {
8         str = ch + str;
9     }
10    return str;
11 }
```

<https://qz.com/646467/how-one-programmer-broke-the-internet-by-deleting-a-tiny-piece-of-code/>



Creating sustainable open source communities is hard

Maybe even harder today than ever before
... because of how open source has changed



Today: more problems than solutions

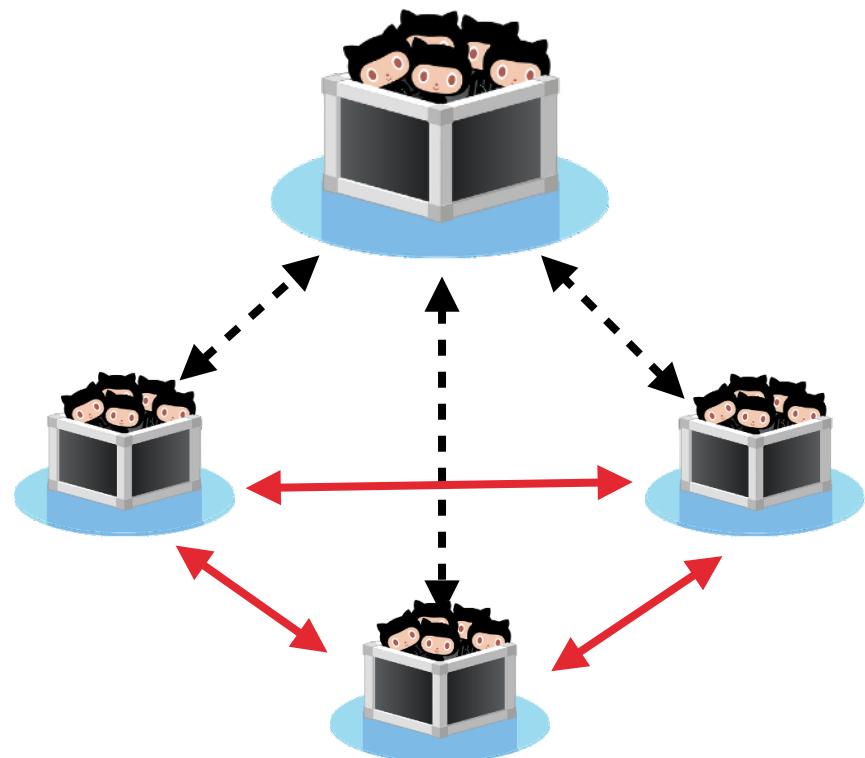


nanook_of_the_burgh

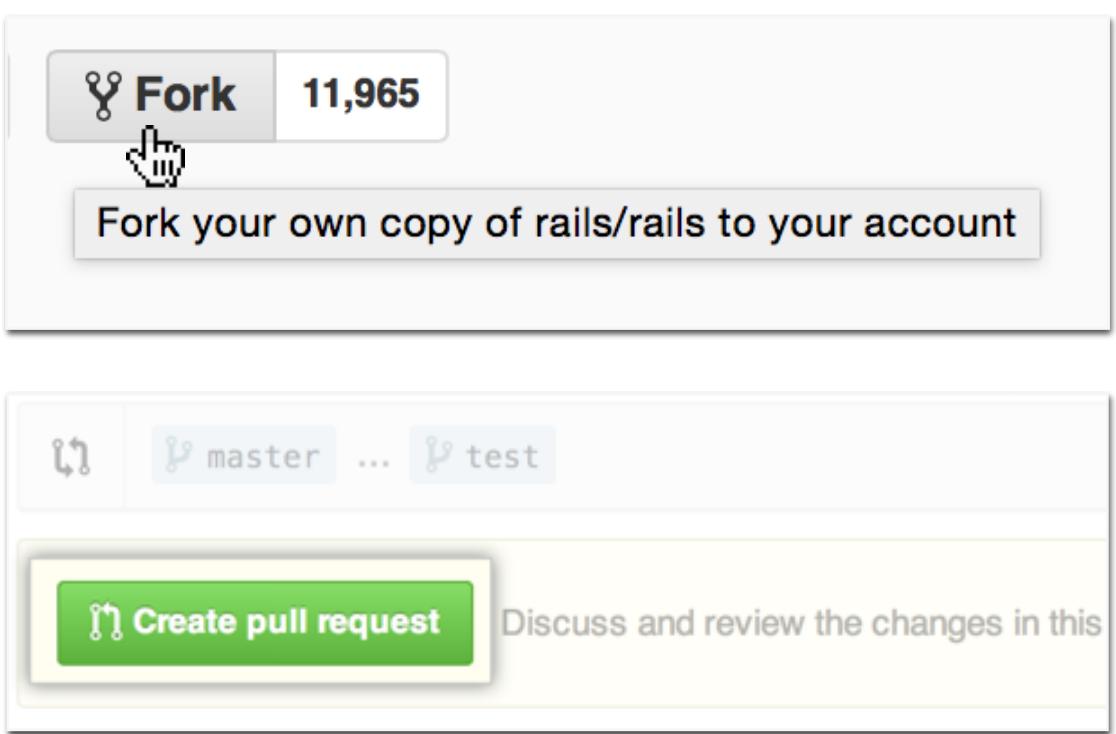
How has open source changed?

Change #1: GitHub standardized the practices

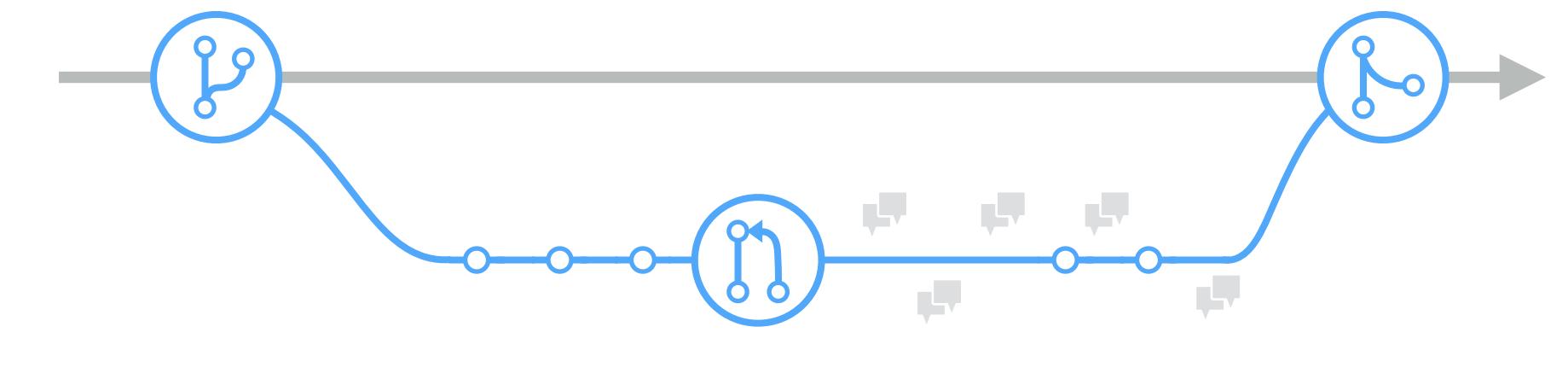
- Git version control



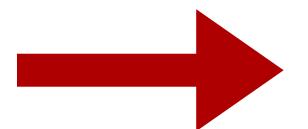
- GitHub UI



- The Pull Request model



- Lower barrier to entry
- Easier to contribute



More production

Change #2: More open source now than ever before

- Explosion of production in the past seven years

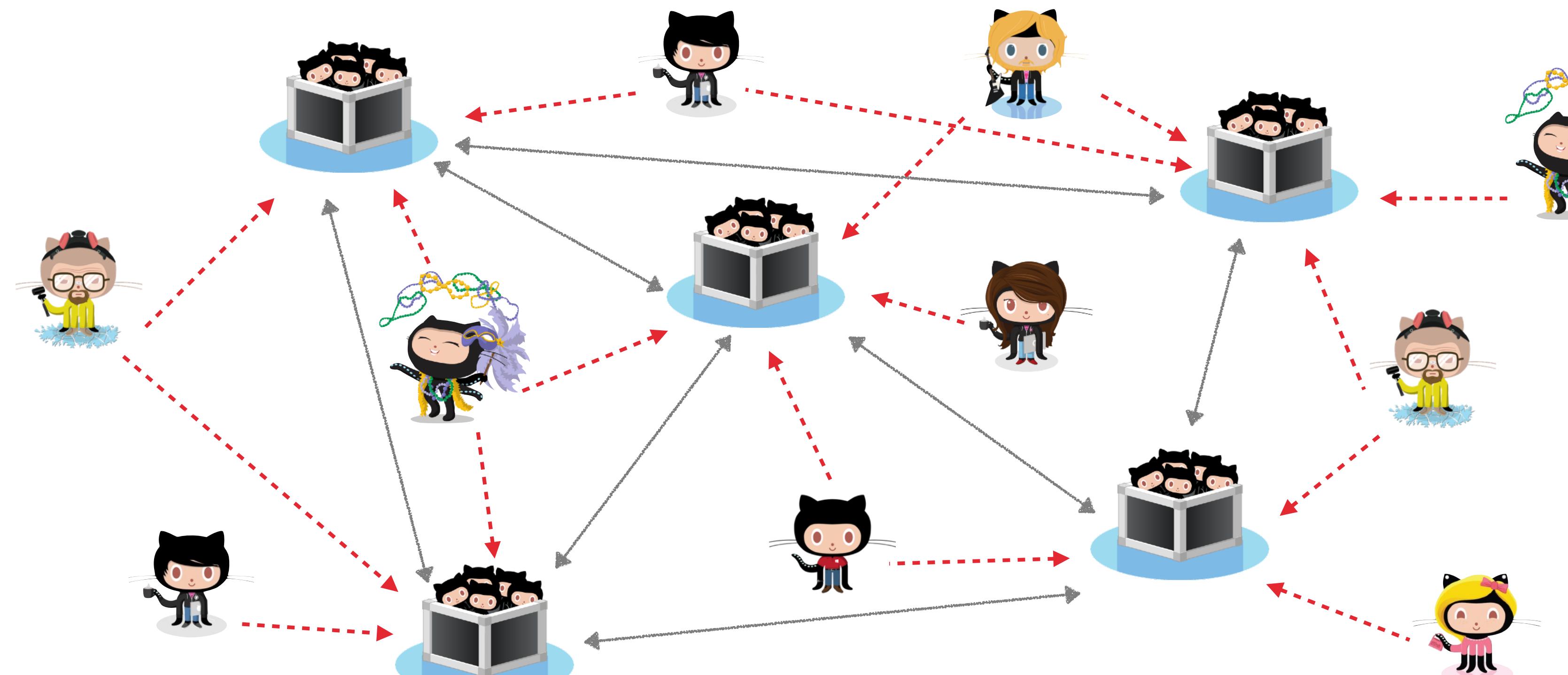


Change #3: High level of transparency

- Profile pages for users and projects
- Rich inferences about people's expertise and level of commitment
- Impacts collaboration, but also recruiting and hiring
 - (Dabbish et al. 2012), (Marlow et al. 2013), (Marlow and Dabbish 2013)

The image shows a composite view of GitHub features. At the top, a GitHub user profile for 'caolan' is displayed, featuring a cartoon cat holding a CV. Below this, a list of 'Popular repositories' is shown, including 'breakfast-repo', 'x86-kernel', and 'jsconf-2015-deck'. To the right, a sidebar lists 'Repositories contributed to' such as 'npm/docs', 'mozilla/publish.webmaker.org', and 'npm/marky-markdown'. The bottom half of the image shows a detailed project page for 'caolan / async'. The page title is 'caolan / async' with a subtitle 'Async utilities for node and the browser'. It includes tabs for Code, Issues (21), Pull requests (6), Projects (0), Wiki, and Insights. Key statistics are listed: 1,629 commits, 11 branches, 72 releases, 206 contributors, and an MIT license. A large image of the 'async' logo is prominently displayed. Below the logo, build status (passing), npm version (v2.6.0), coverage (99%), and other links like Gitter and jsDelivr are shown. A descriptive text block at the bottom states: 'Async is a utility module which provides straight-forward, powerful functions for working with asynchronous JavaScript. Although originally designed for use with Node.js and installable via npm install --save async, it can also be used directly in the browser.'

Change #4: Complex socio-technical ecosystems



Interconnections between people and projects

Can be brittle

NPM ERR!

How one programmer broke the internet by deleting a tiny piece of code

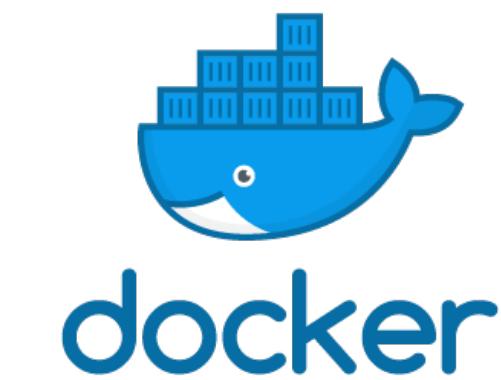
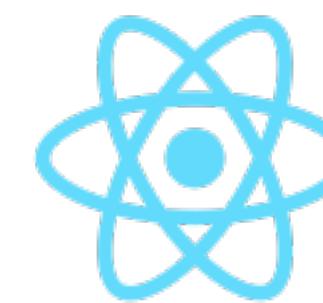
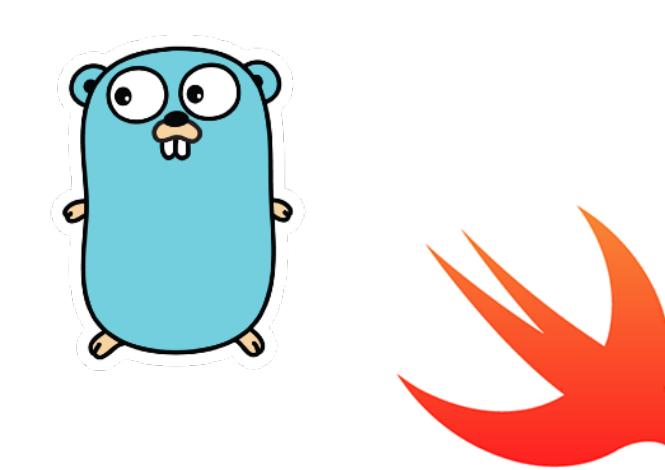
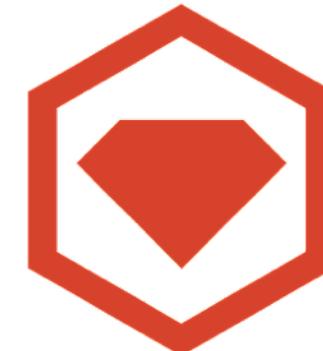
By Keith Collins • March 27, 2016

```
leftpad.js package.json
1 module.exports = leftpad;
2 function leftpad(str, len, ch) {
3     str = String(str);
4     var i = -1;
5     if (!ch && ch !== 0) ch = ' ';
6     len = len - str.length;
7     while (++i < len) {
8         str = ch + str;
9     }
10    return str;
11 }
```

<https://qz.com/646467/how-one-programmer-broke-the-internet-by-deleting-a-tiny-piece-of-code/>

Change #5: Increasing commercialization and professionalization

- Historically
 - Community-based projects (Python, RubyGems, Twisted)
- Currently
 - Lots of commercial involvement
 - Companies (Go - Google, React - Facebook, Swift - Apple)
 - Startups (Docker, npm, Meteor)



- 23% of respondents to 2017 GitHub survey: job duties include contributing to open source

<http://opensourcesurvey.org/2017/>

Change #6: High expectations toward the quality, reliability, and security of open source infrastructure

- Equifax (market cap \$14 billion) built products on top of open-source infrastructure, including Apache Struts
- Equifax did not make any contributions to open source projects
- A flaw in Apache Struts contributed to the breach (CVE-2017-5638)
- Equifax publicly blamed (with national news coverage) Apache Struts for the breach

Equifax confirms Apache Struts security flaw it failed to patch is to blame for hack

The company said the March vulnerability was exploited by hackers.



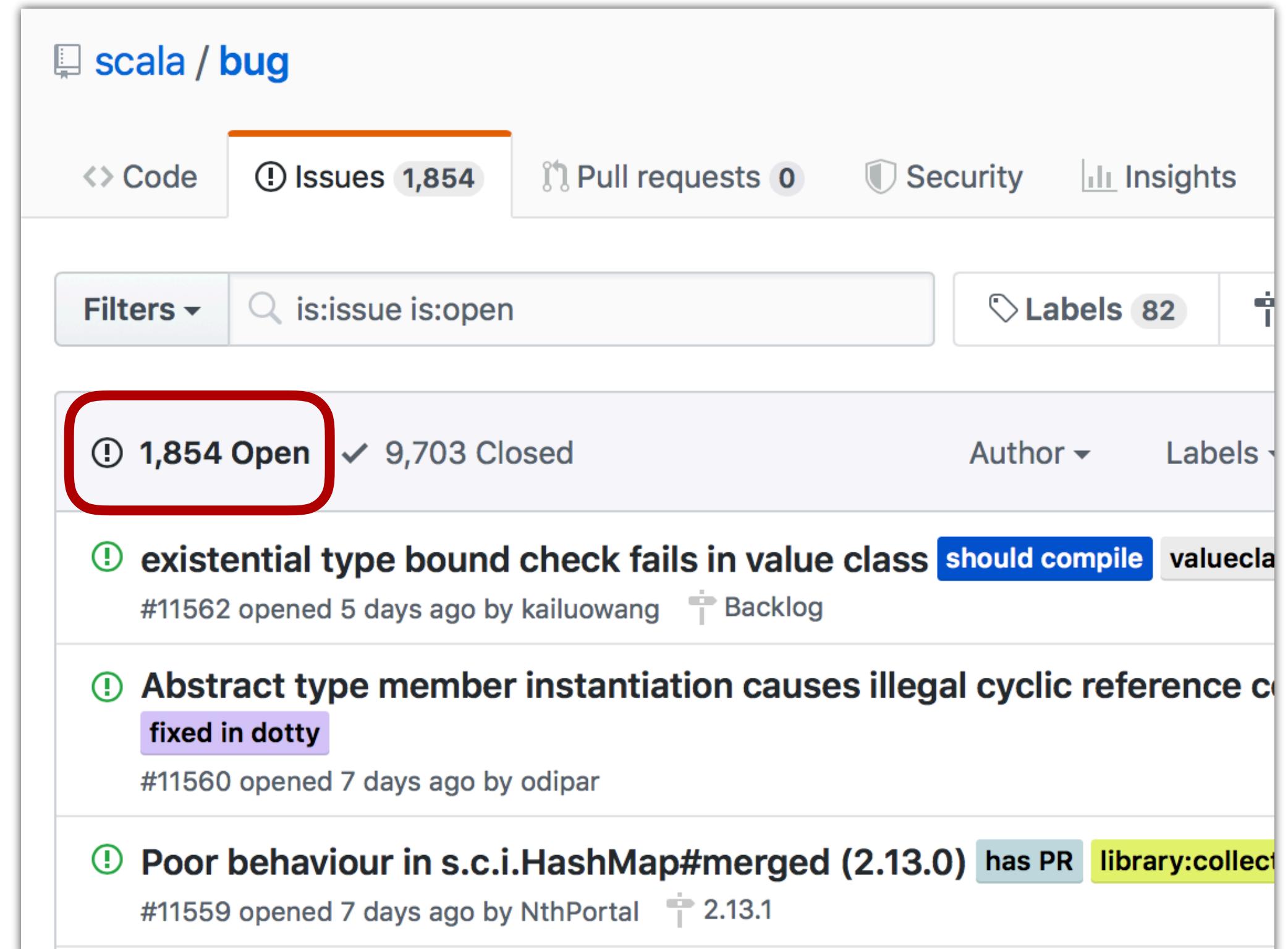
By Zack Whittaker | September 14, 2017 -- 01:27 GMT (18:27 PDT) | Topic: Security



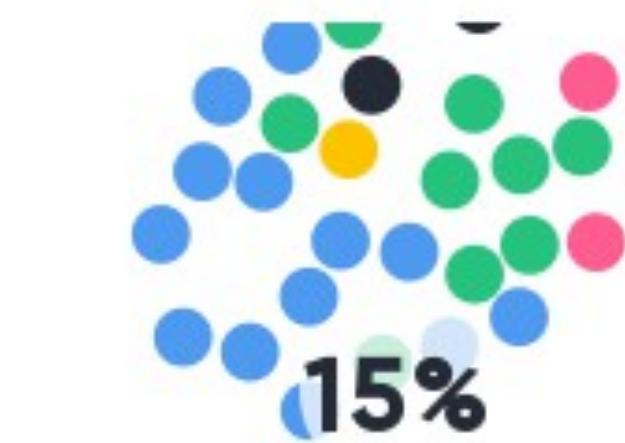
<https://www.zdnet.com/article/equifax-confirms-apache-struts-flaw-it-failed-to-patch-was-to-blame-for-data-breach/>

Change #7: High level of demands & stress

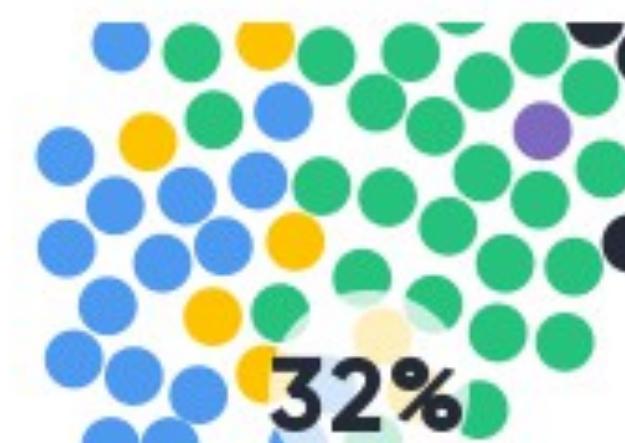
- Easy to report issues / submit PRs
 - Growing volume of requests
- Social pressure to respond quickly
 - Otherwise, off-putting to newcomers
(Steinmacher et al. 2015)
- Entitlement, unreasonable requests from users:
 - *“I have been waiting 2 years for Angular to track the ‘progress’ event and it still can’t get it right?!?!”*
 - *“Thank you for your ever useless explanations.”*



Do you ever feel overwhelmed with the amount of feature requests and bug reports in your open source projects?



Often



Sometimes



Seldom



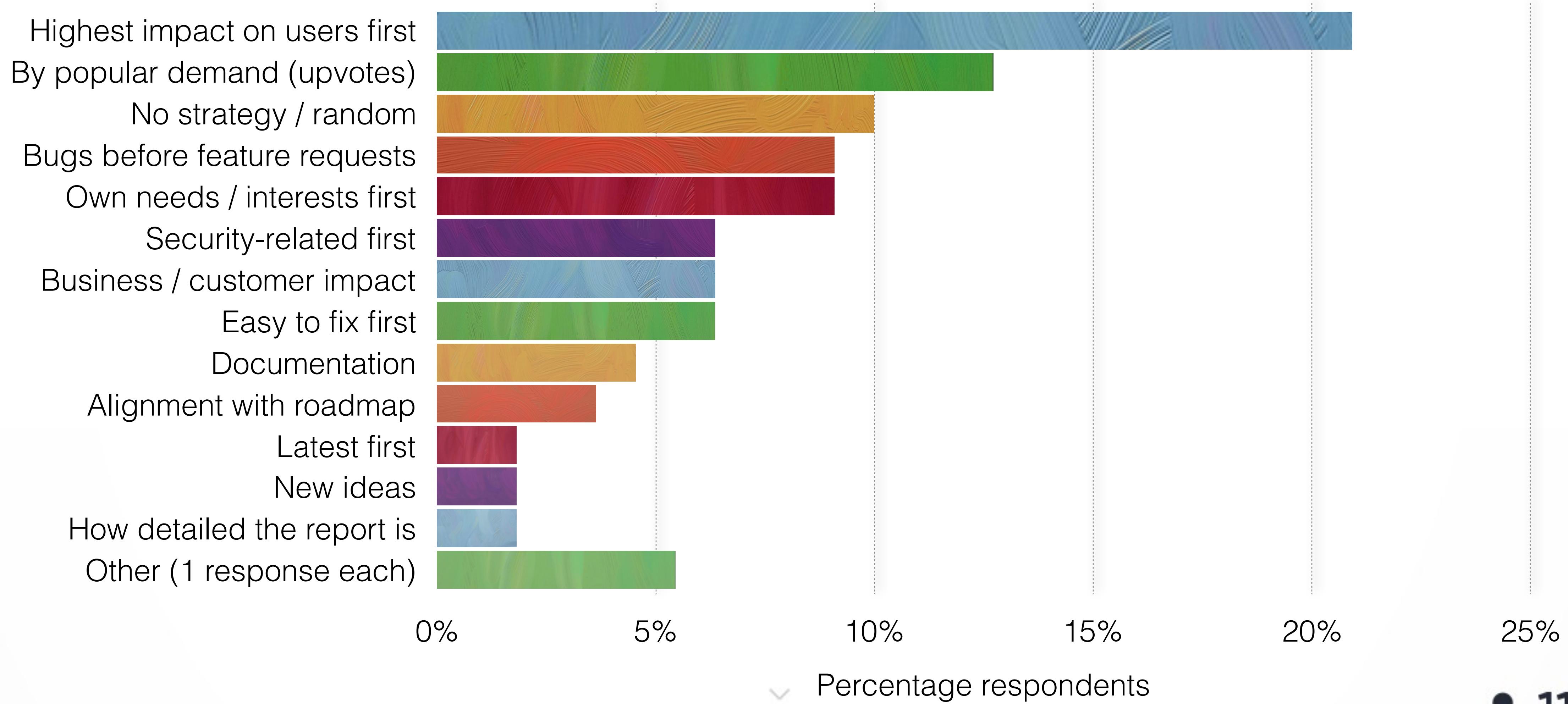
Never

How do you engage with open source?

- I maintain one or more open source projects
- I regularly contribute to open source projects
- I regularly use open source projects in proprietary work
- I rely primarily on closed source
- I don't code
- Unknown

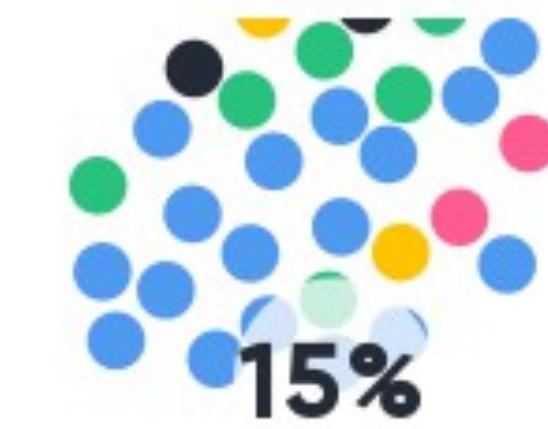
182

How do you prioritize issues / pull requests?

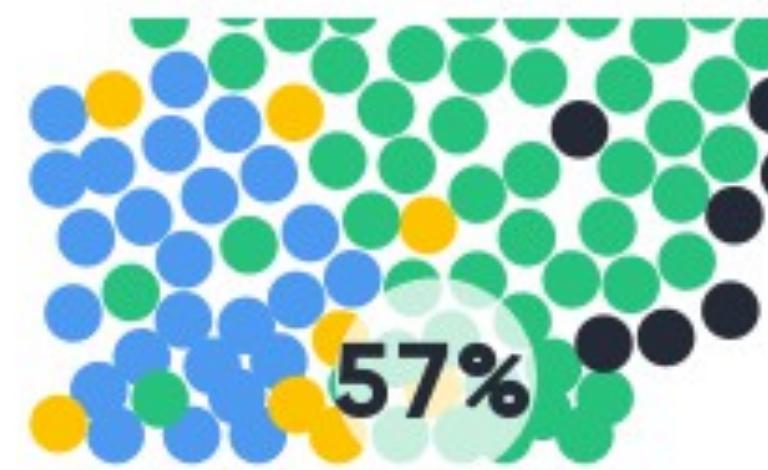


115

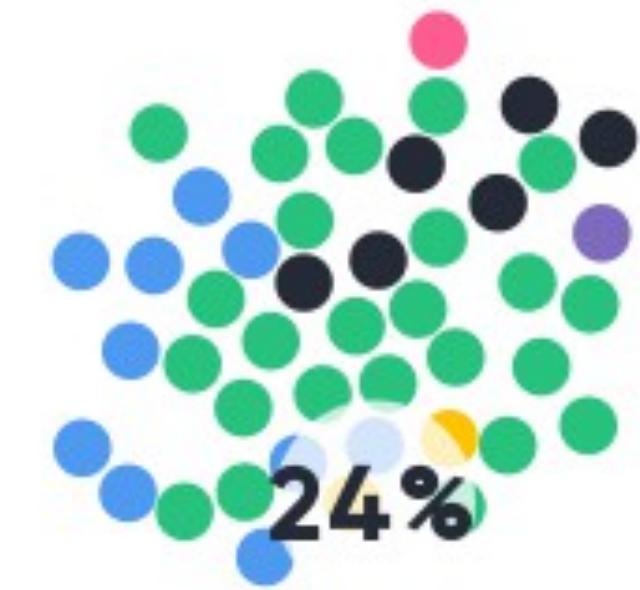
Do you feel the interaction between developers and users of your projects is healthy and sustainable?



Almost Always



To a Considerable Degree



Occasionally



Seldom

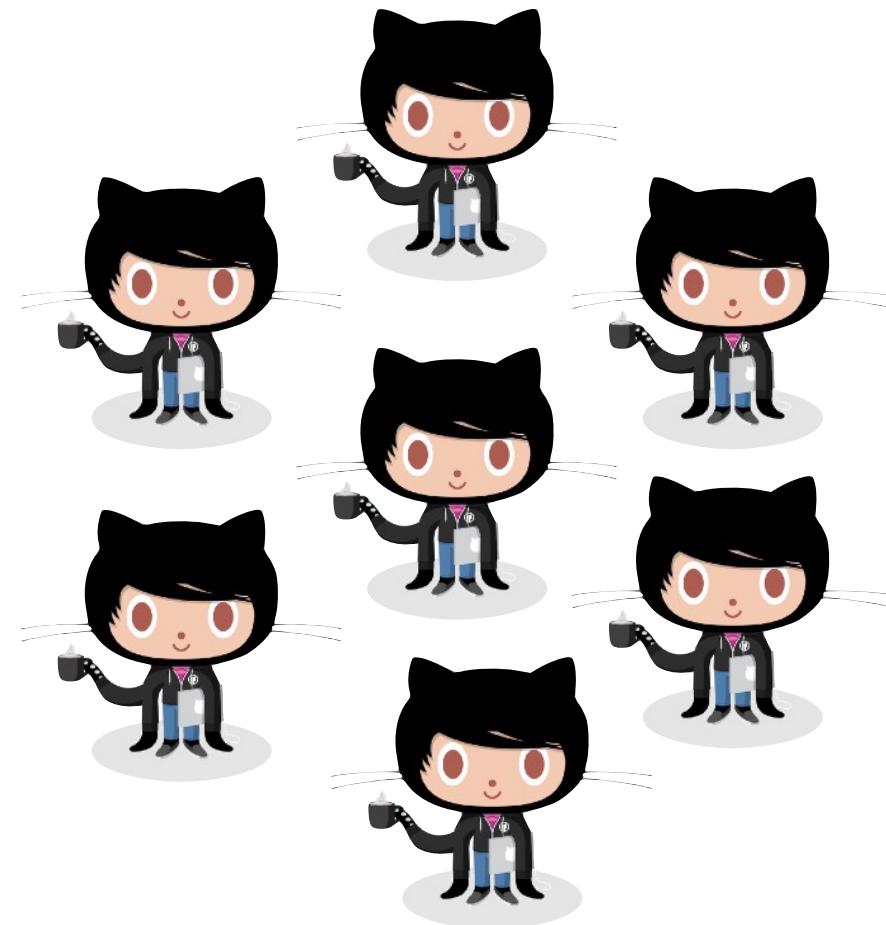
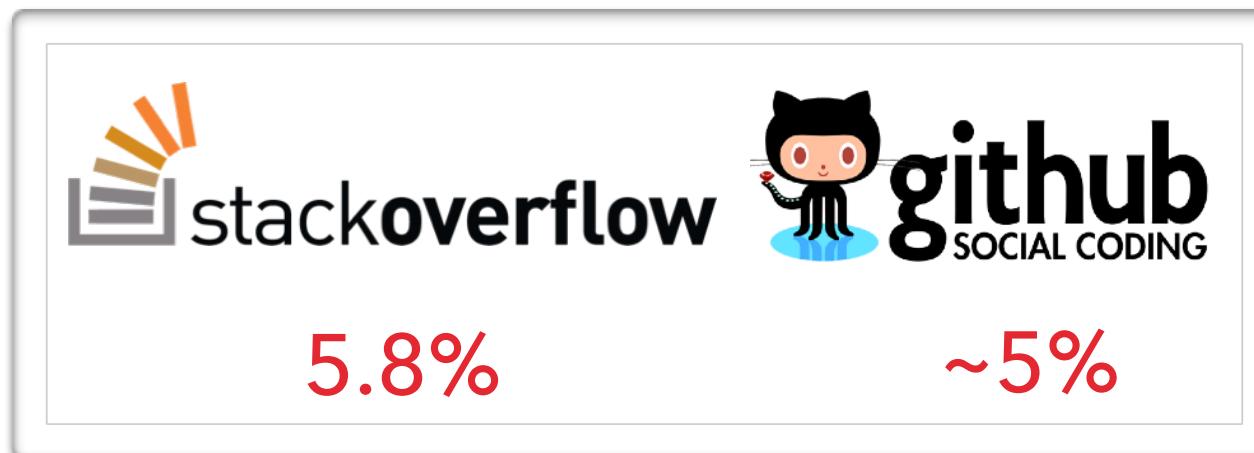
How do you engage with open source?

- I maintain one or more open source projects
- I regularly contribute to open source projects
- I regularly use open source projects in proprietary work
- I rely primarily on closed source
- I don't code
- Unknown

188

Change #8: Low demographic diversity

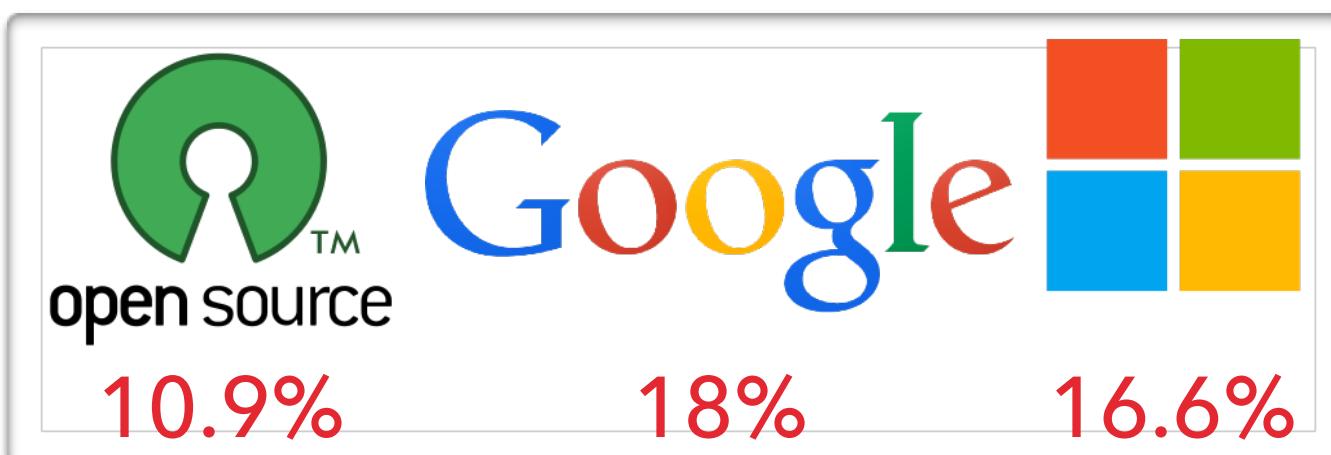
- Gender representation reality



- Expectation



“More about the contributions to the code than the ‘characteristics’ of the person”



“Any demographic identity is irrelevant”

“Code sees no color or gender”

- FLOSS 2013: A survey dataset about free software contributors: challenges for curating, sharing, and combining G Robles, L Arjona-Reina, B Vasilescu, A Serebrenik, JM Gonzalez-Barahona. MSR 2014
- Google Diversity (2015) www.google.com/diversity/index.html#chart
- Inside Microsoft (2015) <https://goo.gl/nT4YiI>

- Exploring the data on gender and GitHub repo ownership Alyssa Frazee. <http://alyssafrazee.com/gender-and-github-code.html>
- Stack Overflow 2015 Developer Survey (26,086 people from 157 countries) <http://stackoverflow.com/research/developer-survey-2015#profile-gender>

- Perceptions of Diversity on GitHub: A User Survey. Vasilescu, B., Filkov, V., and Serebrenik, A. CHASE 2015

Aside: Why should you care about gender diversity?

Productivity boosts

DIVERSE TEAMS ARE MORE PRODUCTIVE!

vs.

Other confounds held fixed, **higher team diversity (gender & tenure)** is associated with **increased code production** (commits per quarter).

But small effects!

Inclusivity helps everyone

Why care? Inclusivity

- Reduces social isolation → unique experience → unique perspective
- Reduces economic costs to health care system

s n

© Anita Sarma & Margaret Burnett, Oregon State U

- Gender and tenure diversity in GitHub teams. Vasilescu, B., Posnett, D., Ray, B., Brand, M.G.J. van den, Serebrenik, A., Devanbu, P., and Filkov, V. *CHI 2015*

What have we learned through empirical research?

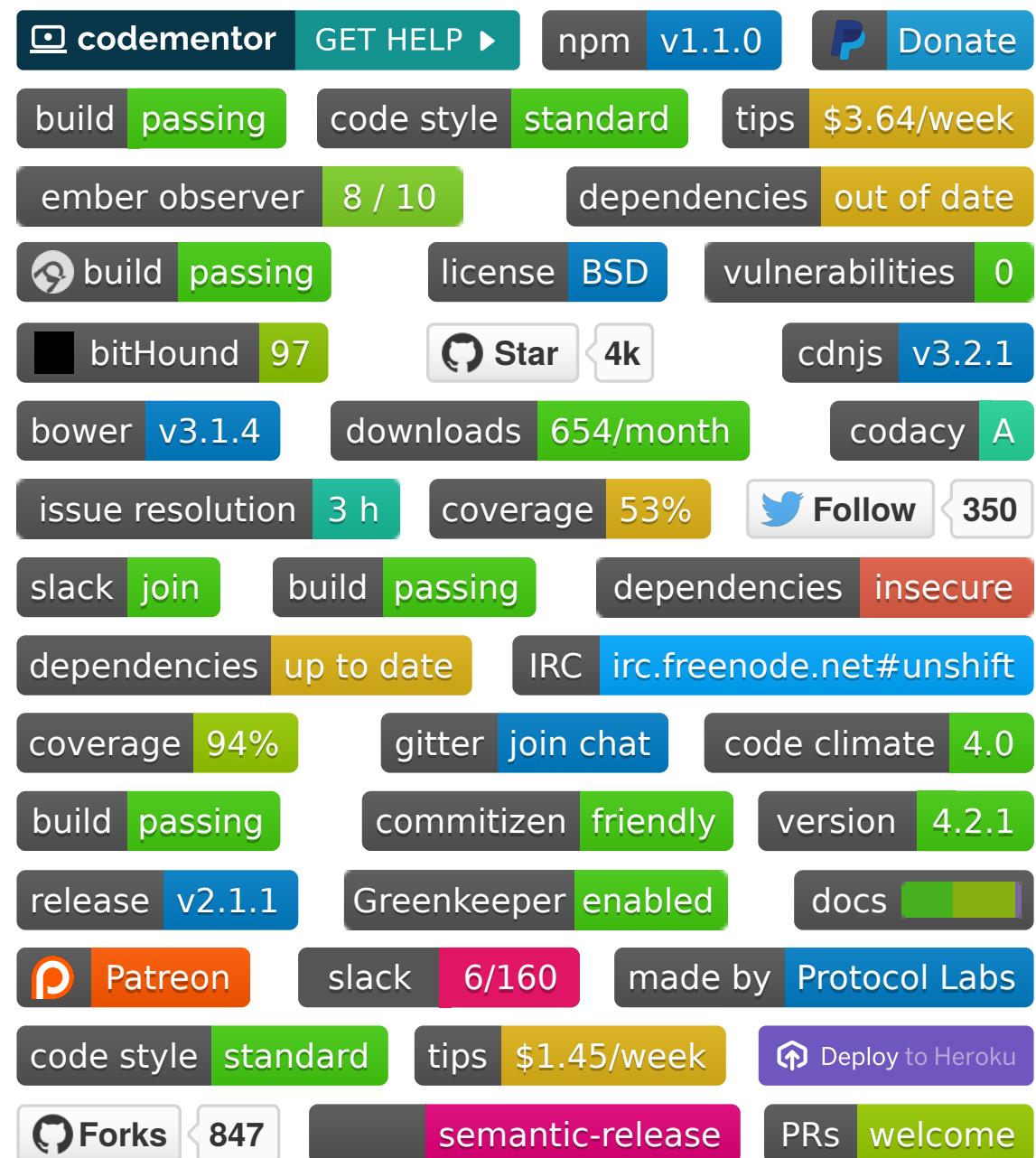
Almost everything is archived.

Data can be mined & analyzed.

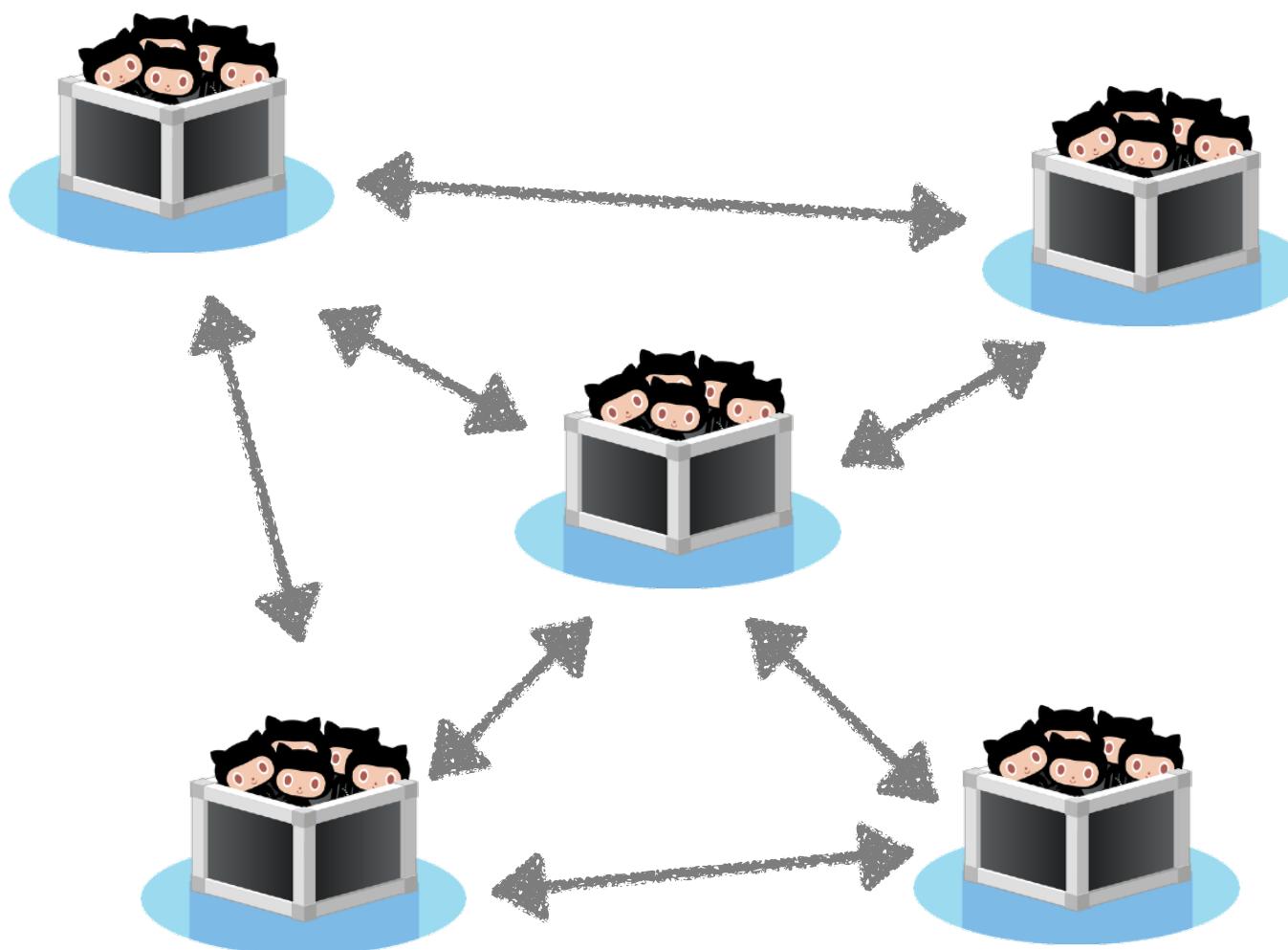
- Understand the effects of these changes
- Reduce / reverse the negative effects

Three examples

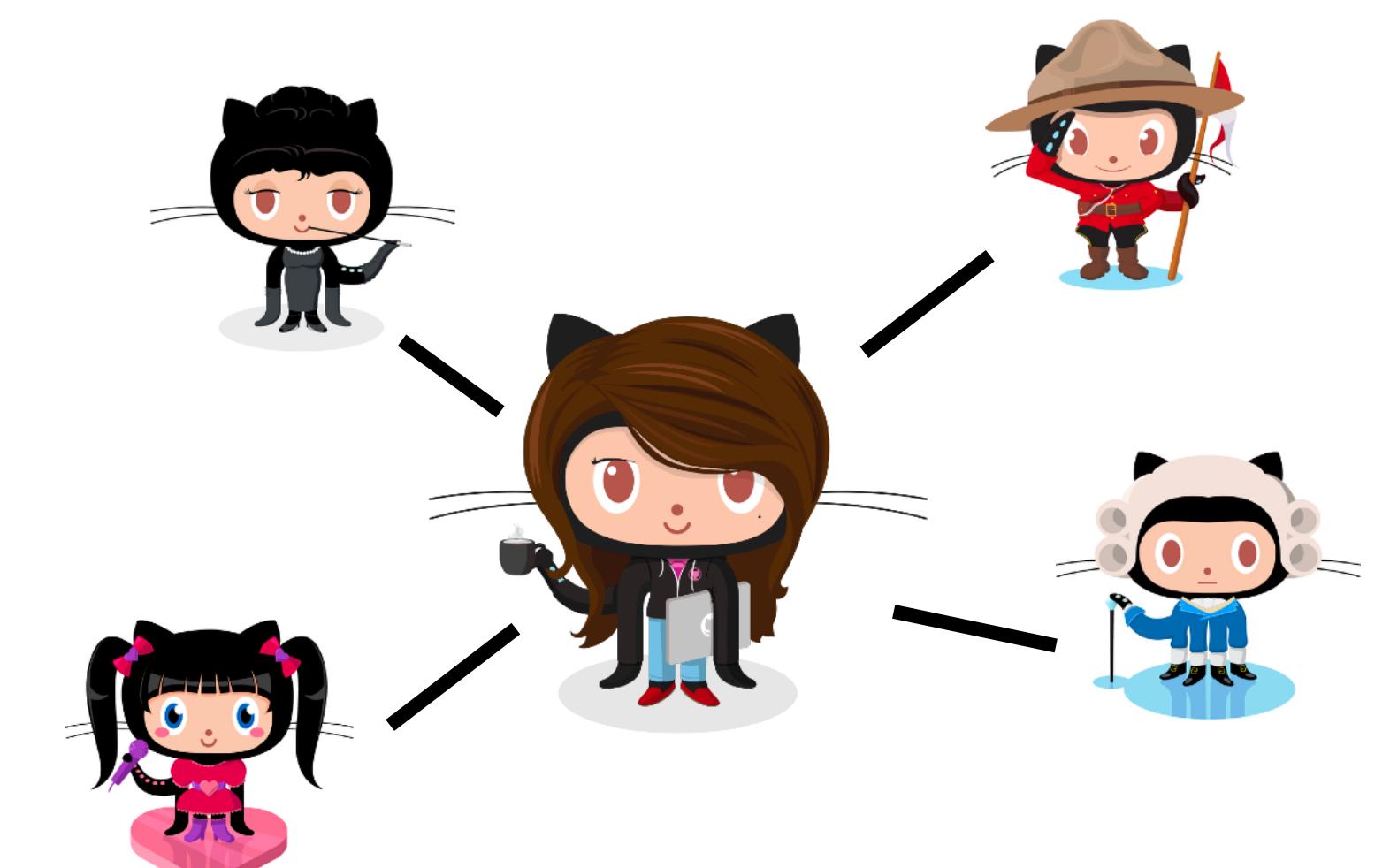
Leveraging transparency



Considering the whole ecosystem



Building social capital

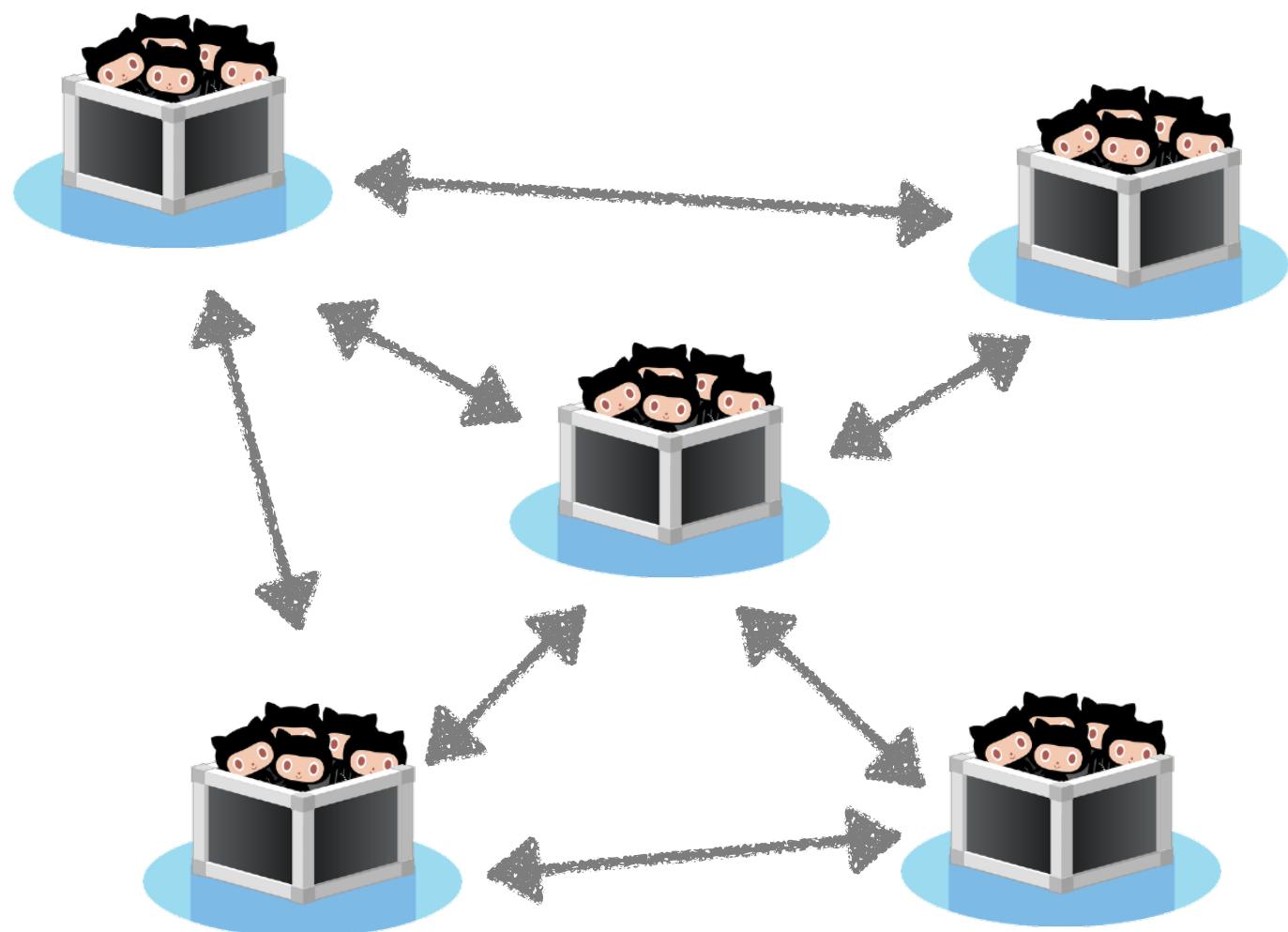


Three examples

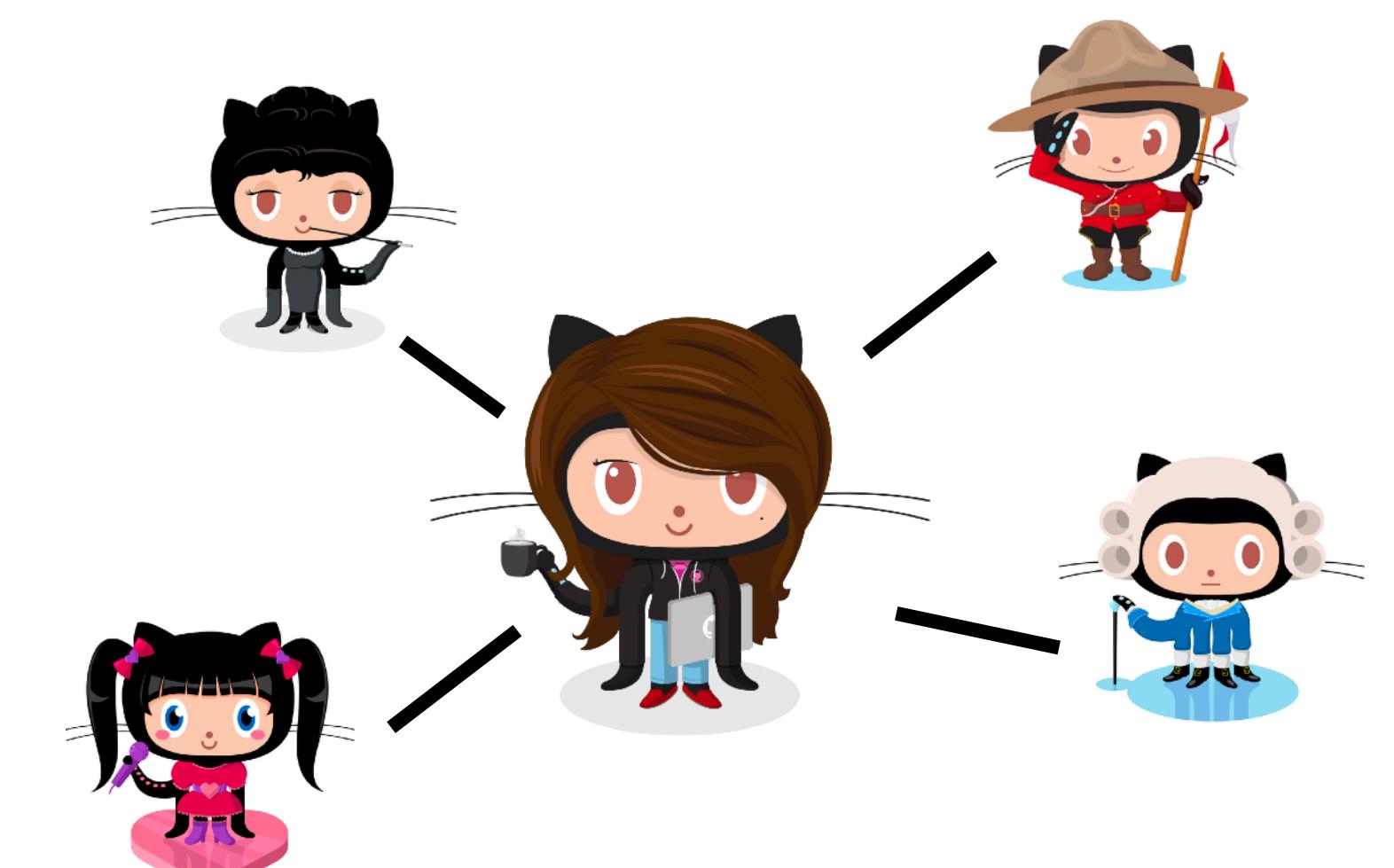
Leveraging transparency



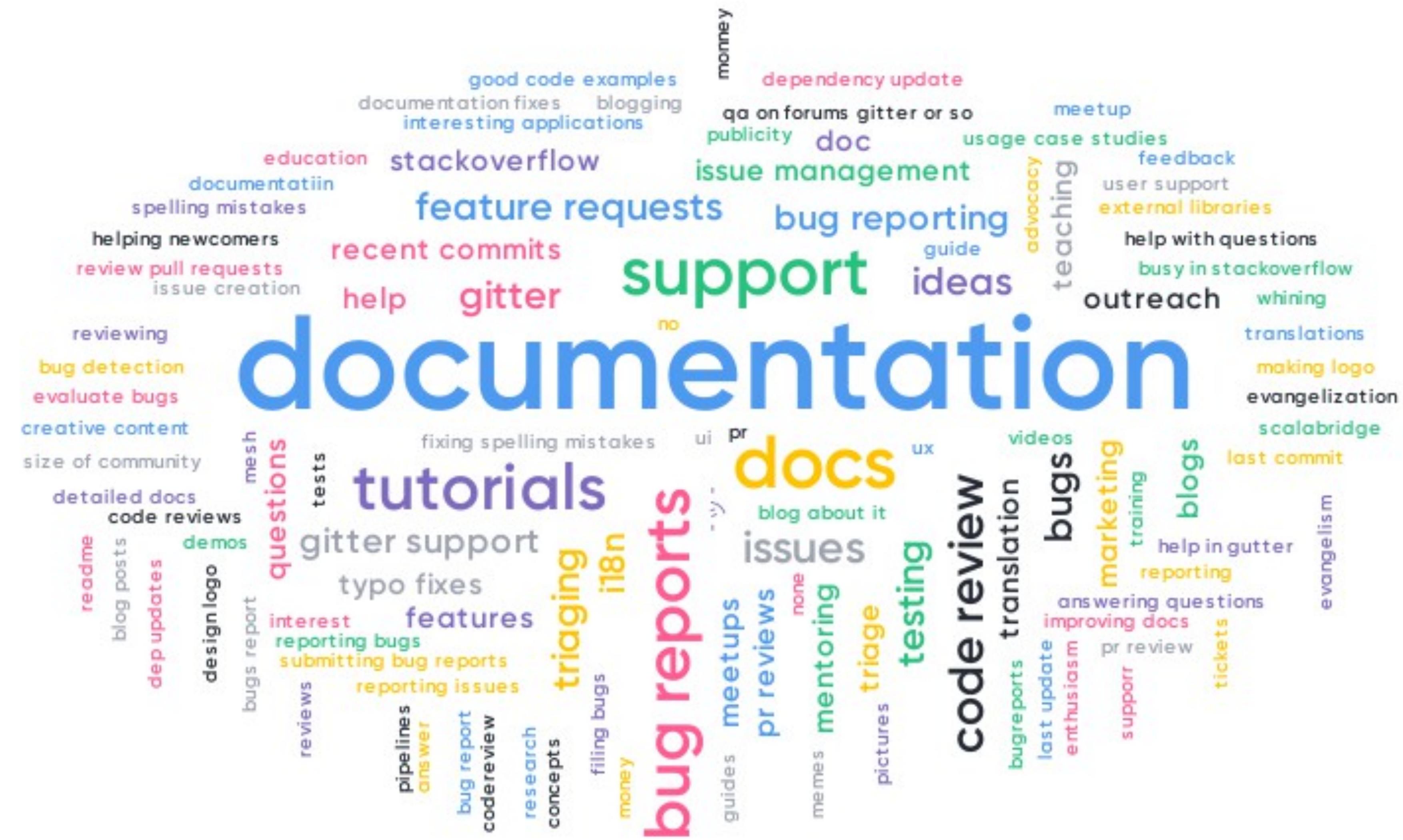
Considering the whole ecosystem



Building social capital



What non-code contributions do you do / see others doing in your open-source projects?



172

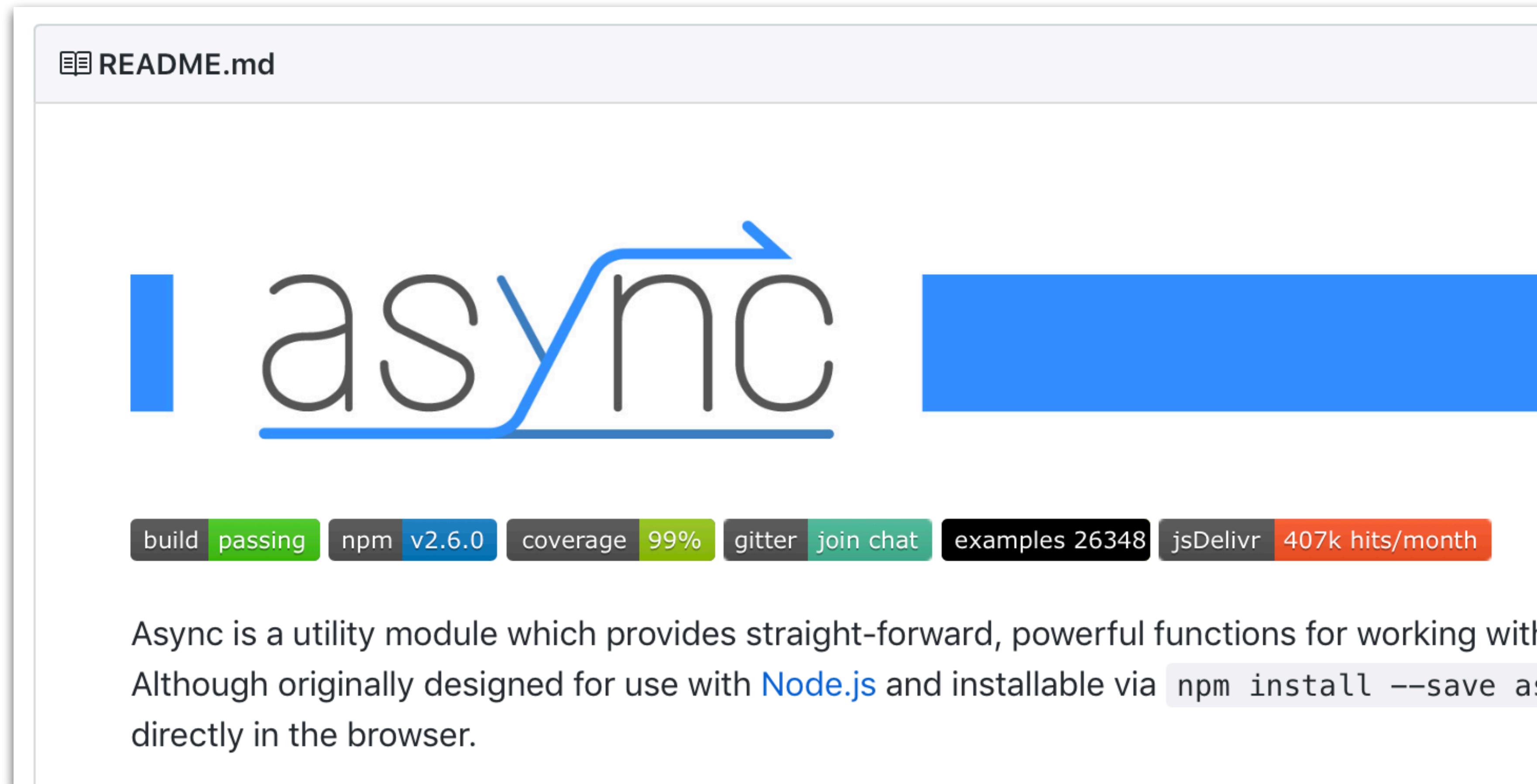
Transparency is already a defining characteristic of the environment

This screenshot shows a GitHub profile page. At the top right, there are buttons for Contributions, Repositories, Public activity, Follow, and a dropdown menu. Below this, the profile picture is a cartoon cat holding a laptop labeled 'CV'. The 'Popular repositories' section lists five repos: breakfast-repo, x86-kernel, jsconf-2015-deck, and ratpack. The 'Repositories contributed to' section lists npm/docs, mozilla/publish.webmaker.org, npm/marky-markdown, artisan-tattoo/assistant-frontend, and npm/npm-camp. The 'Public contributions' section includes a heatmap of activity over time and a summary: Contributions in the last year: 1,886 total (Jan 24, 2015 – Jan 24, 2016), Longest streak: 37 days (October 7 – November 12), Current streak: 7 days (January 18 – January 24). On the left sidebar, it says 'npm, inc' and shows icons for location, mail, and code. It also indicates 'Joined on Oct 31, 2011'. At the bottom, it shows 776 Followers, 38 Starred, and 15 Following. The 'Organizations' section lists four organizations with their respective logos.

This screenshot shows the GitHub repository page for 'caolan / async'. At the top right, it shows Watch (721), Star (23,937), Fork (2,203), and Insights. The repository title is 'caolan / async'. Below it, tabs for Code, Issues (21), Pull requests (6), Projects (0), Wiki, and Insights are visible. The description reads: 'Async utilities for node and the browser <http://caolan.github.io/async/>'. The repository stats are: 1,629 commits, 11 branches, 72 releases, 206 contributors, and MIT license. A large 'async' logo is prominently displayed. Below the logo, social links and metrics are listed: build passing, npm v2.6.0, coverage 99%, gitter, join chat, examples 26348, jsDelivr, and 407k hits/month. The text below the logo states: 'Async is a utility module which provides straight-forward, powerful functions for working with [asynchronous JavaScript](#). Although originally designed for use with [Node.js](#) and installable via `npm install --save async`, it can also be used directly in the browser.'

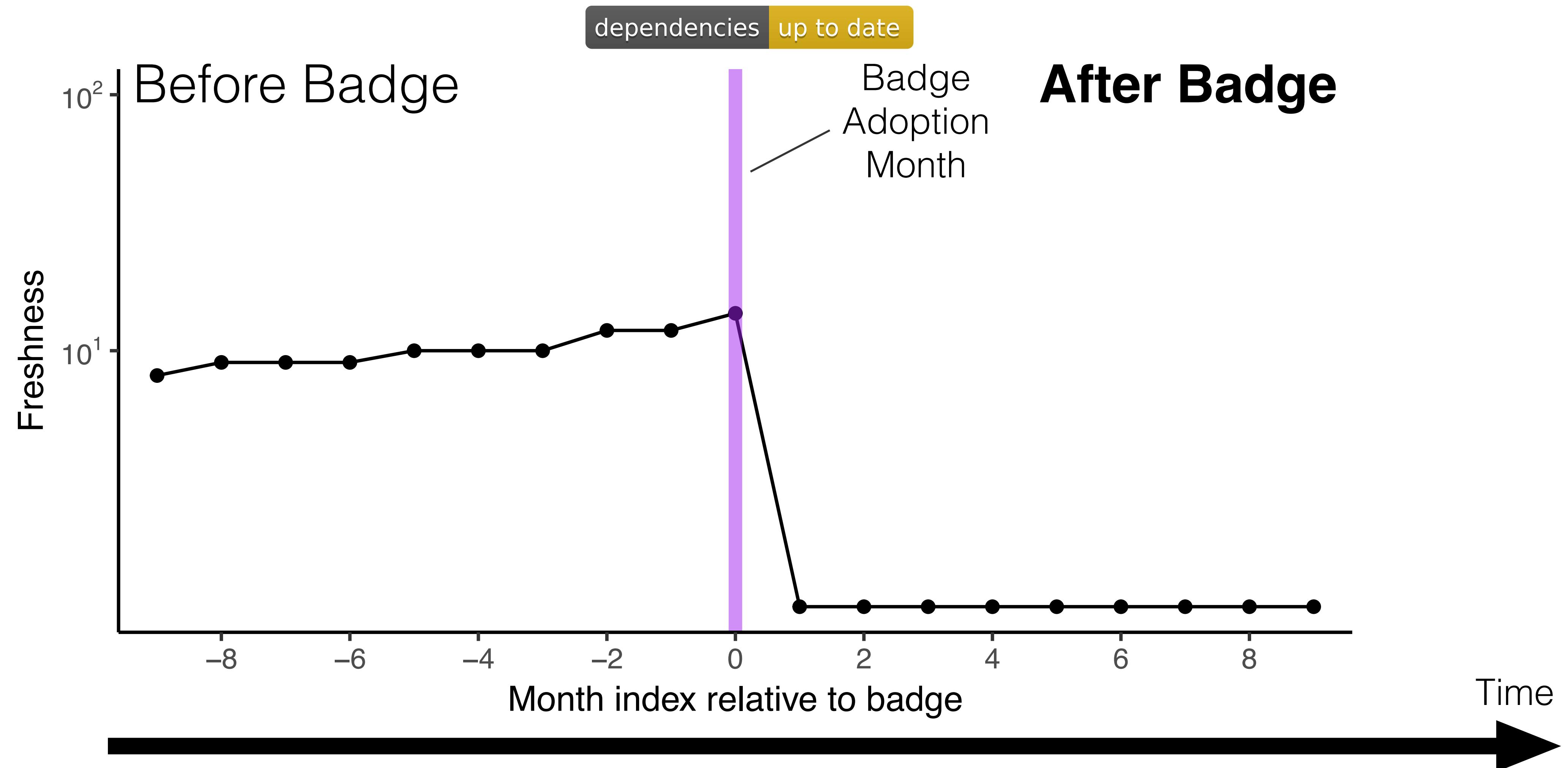
Signals are customizable

- E.g., repository badges

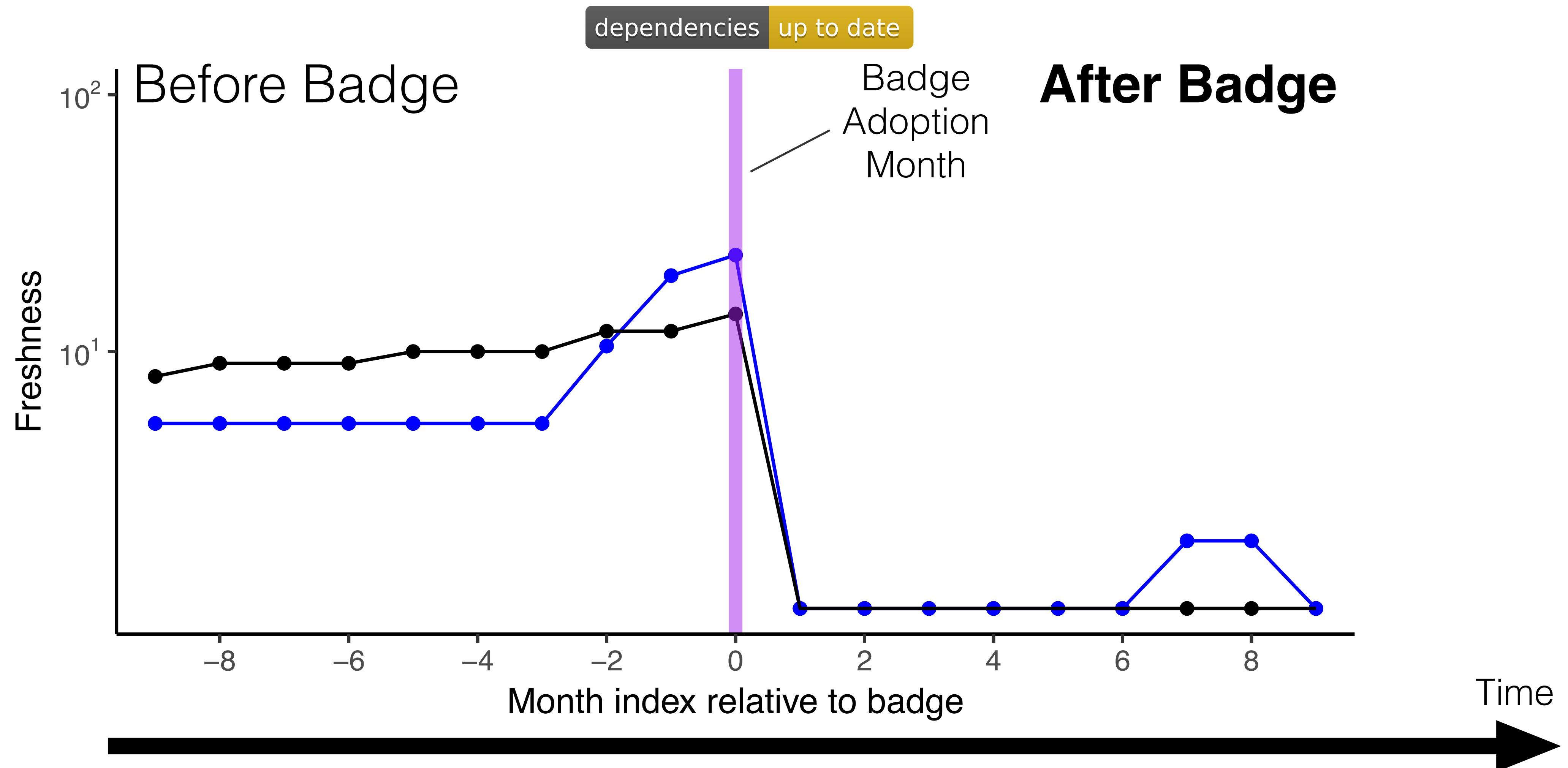


- Adding Sparkle to Social Coding: An Empirical Study of Repository Badges in the npm Ecosystem. Trockman, A., Zhou, S., Kästner, C., and Vasilescu, B. ICSE 2018

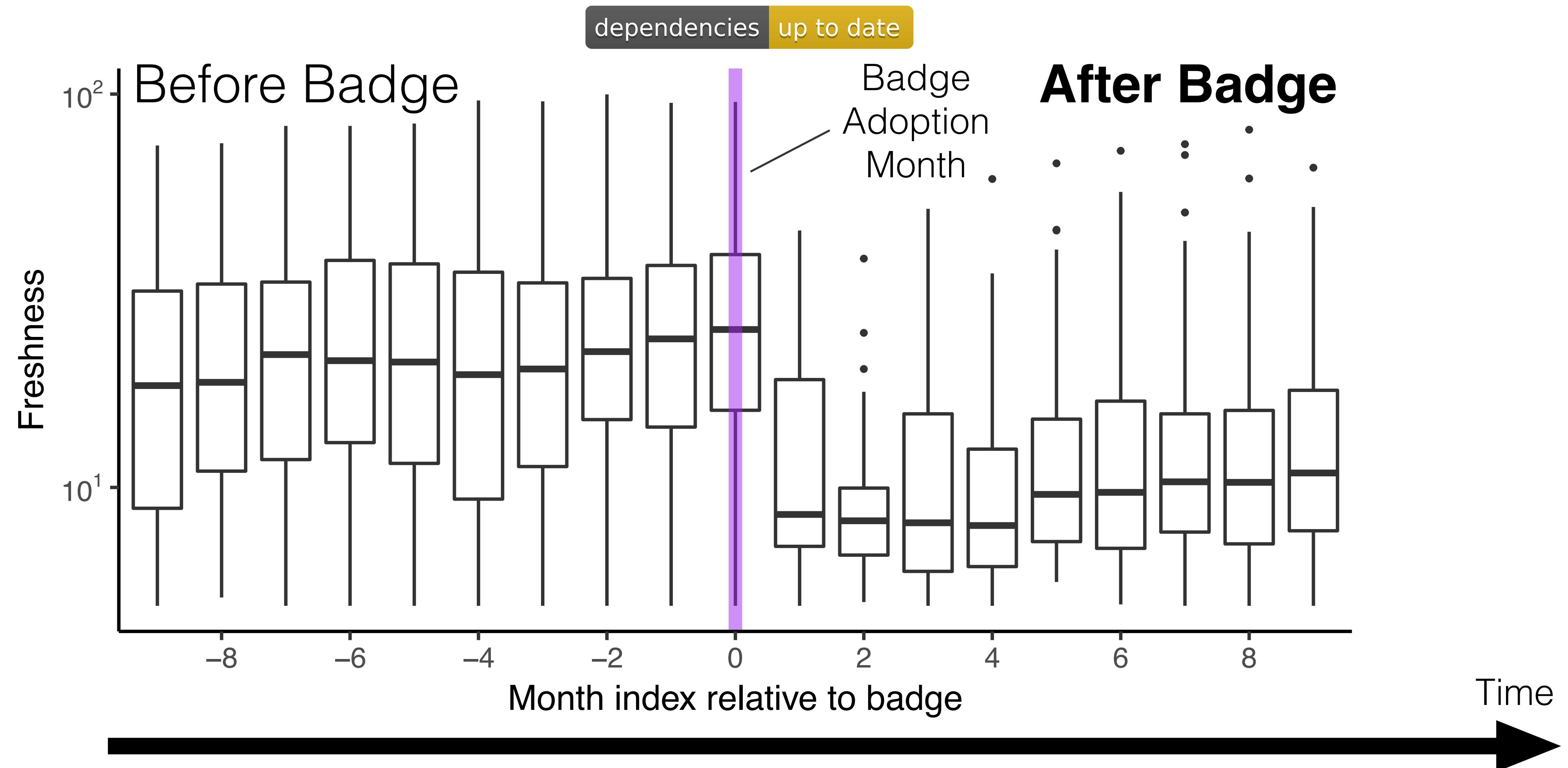
Time Series Analysis



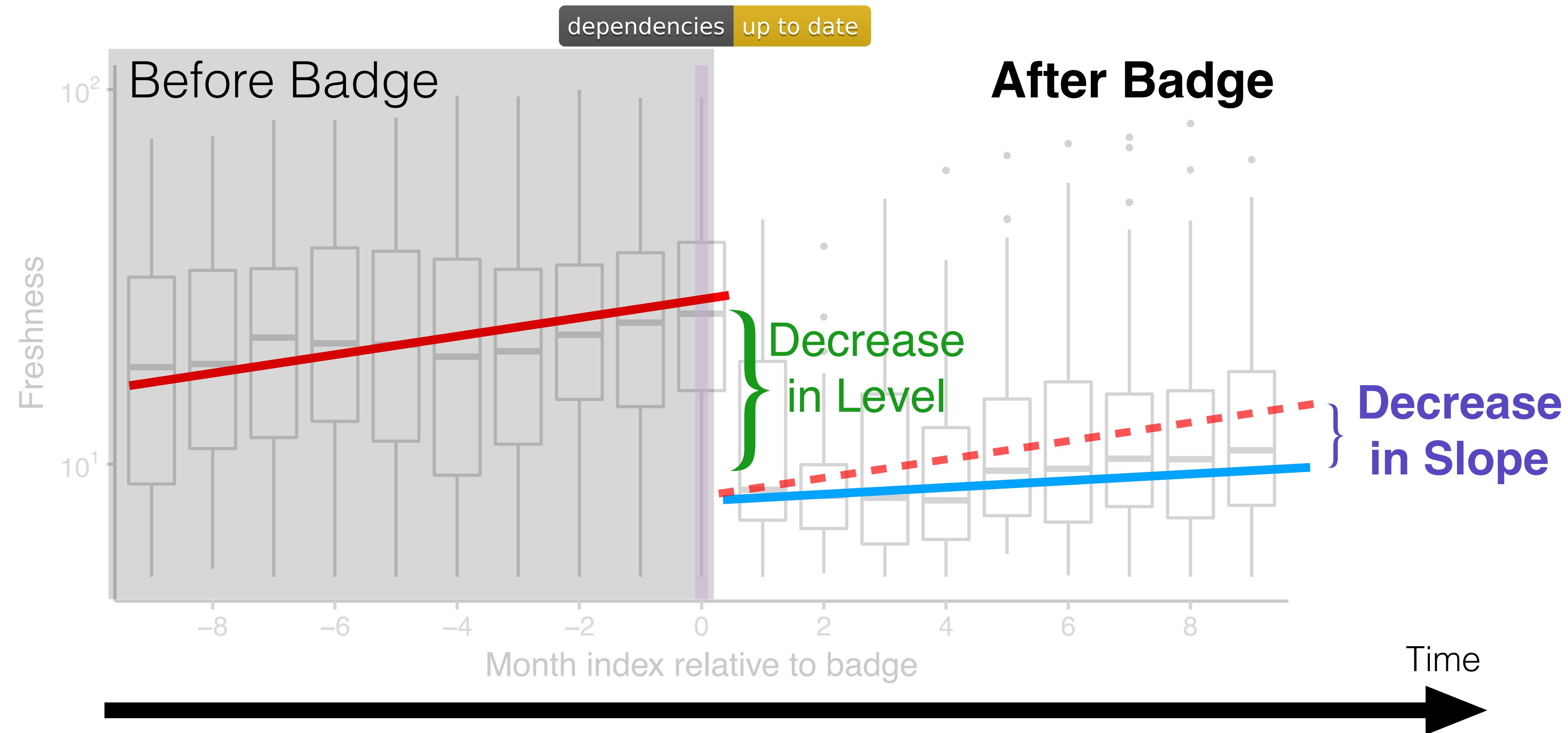
Time Series Analysis



Time Series Analysis



Time Series Analysis



Badges are Reliable Signals

Mostly



Scala badges?

The Scala programming language <http://www.scala-lang.org/>

scala scala-compiler scala-programming-language scala-library jvm-languages

33,064 commits 9 branches 133 releases 436 contributors Apache-2.0

Branch: 2.13.x New pull request Create new file Upload files Find File Clone or download

SethTisue Merge pull request #8144 from SethTisue/next-will-be-2.13.1 ... Latest commit 8401fd3 2 days ago

admin follow-up to sbt 1.x changes 9 months ago

doc Merge commit 'dcd730fe81' into merge/2.12.x-to-2.13.x-20190205 4 months ago

project Upgrade Dotty to 0.15.0-RC1 20 days ago

scripts sbt 1.2.8 (was 1.2.7) 4 months ago

Scala CI

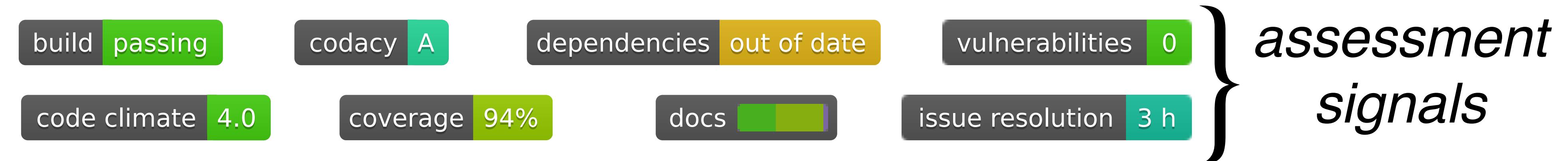
build passing

Once you submit a PR your commits will be automatically tested by the Scala CI.

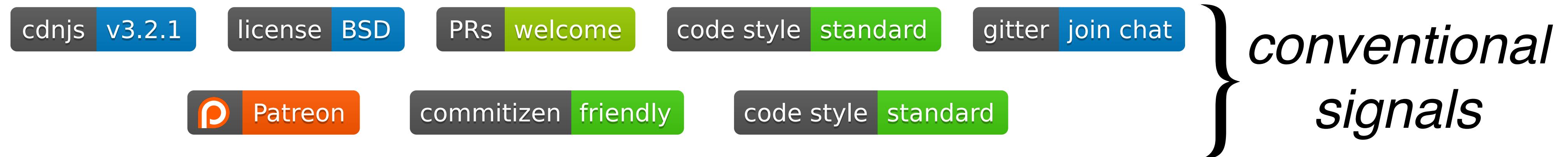
If you see a spurious build failure, you can post /rebuild as a PR comment. The [scabot README](#) lists all available commands.

Take-away: Prefer “assessment” badges

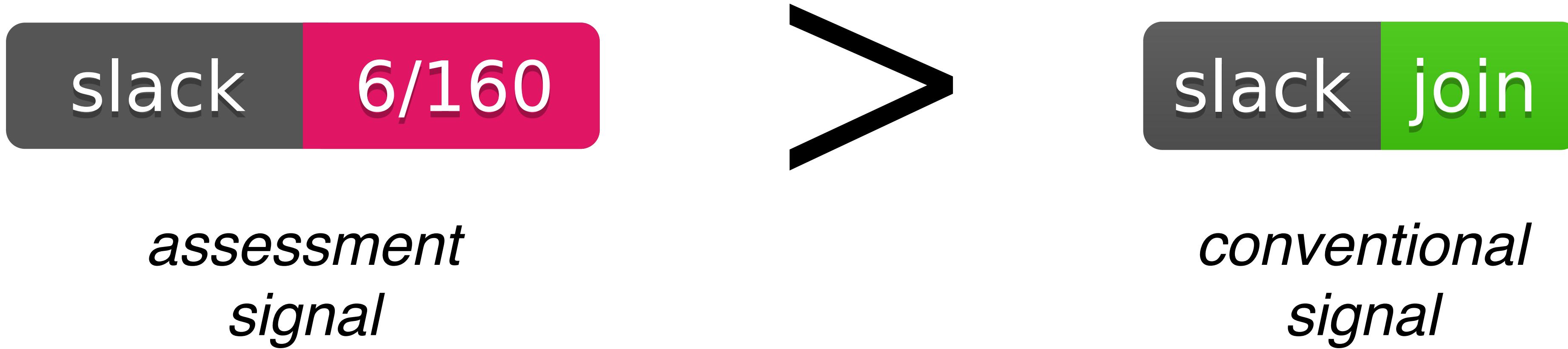
Badges with underlying analyses:



are **stronger predictors** than badges that merely state intentions or provide links:

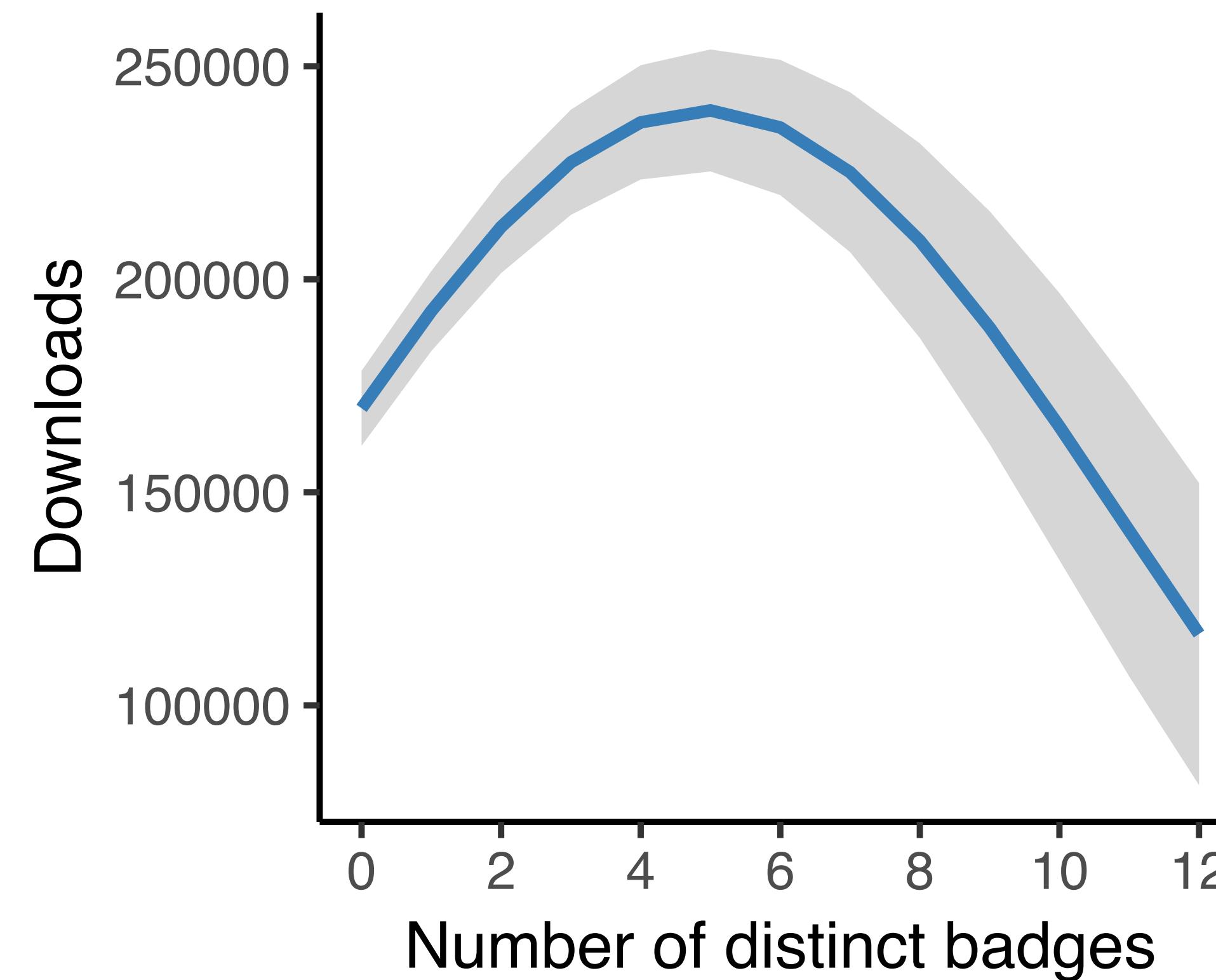


Take-away: Prefer “assessment” badges



Take-away: Don't add too many

Attractiveness wears off beyond 5 badges



“It’s most important that the people seem nice”

How do people choose which project to contribute to?

The **tone of the community** is an important factor in both interviews and model.

maintainers polite ?

Asking for help explicitly is an important factor in the interviews.

PRs welcome help wanted ?

Interviews:

15 GitHub users

Data:

~10K npm packages

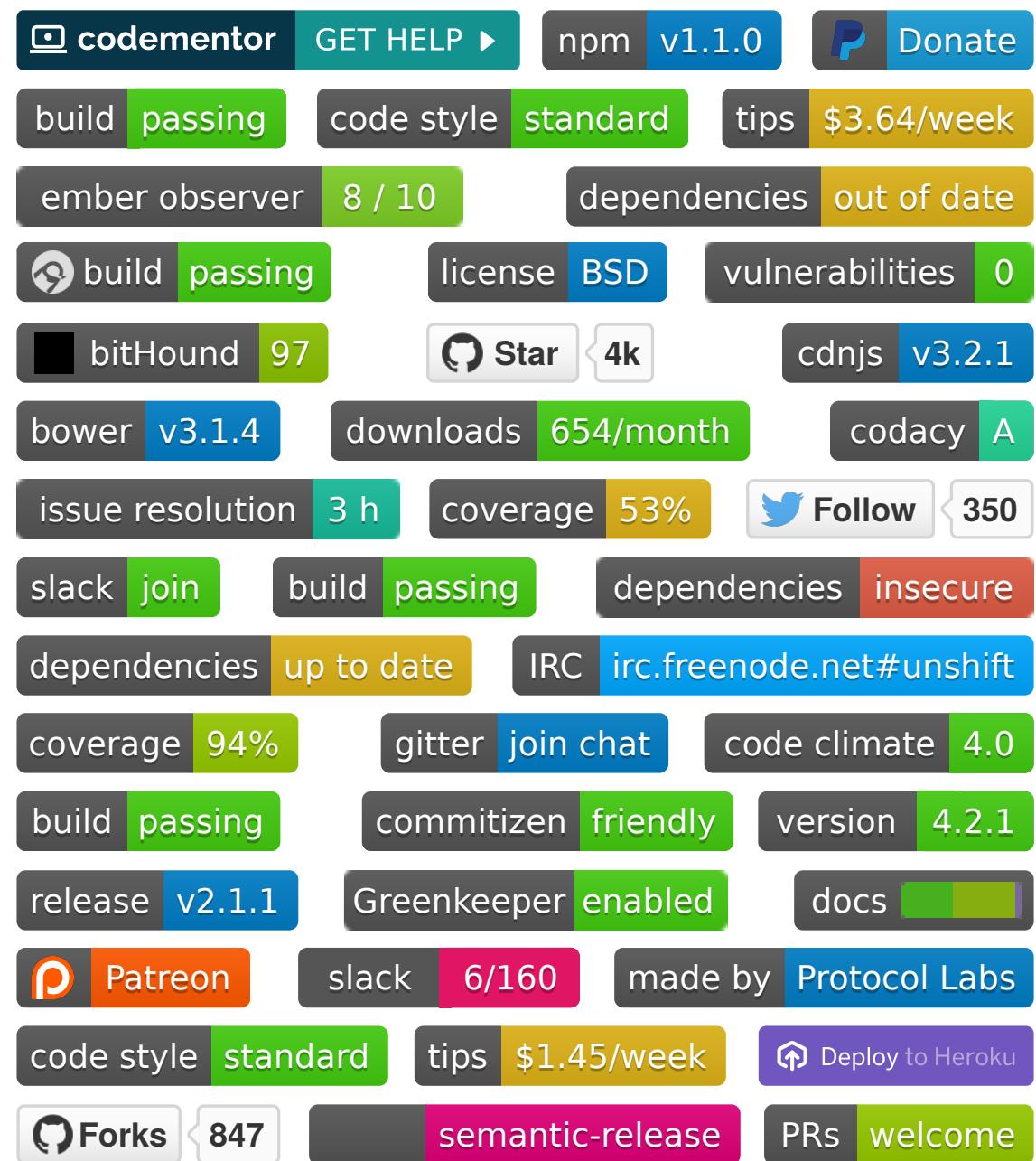
Model:

Logistic regression
(has new contributors)

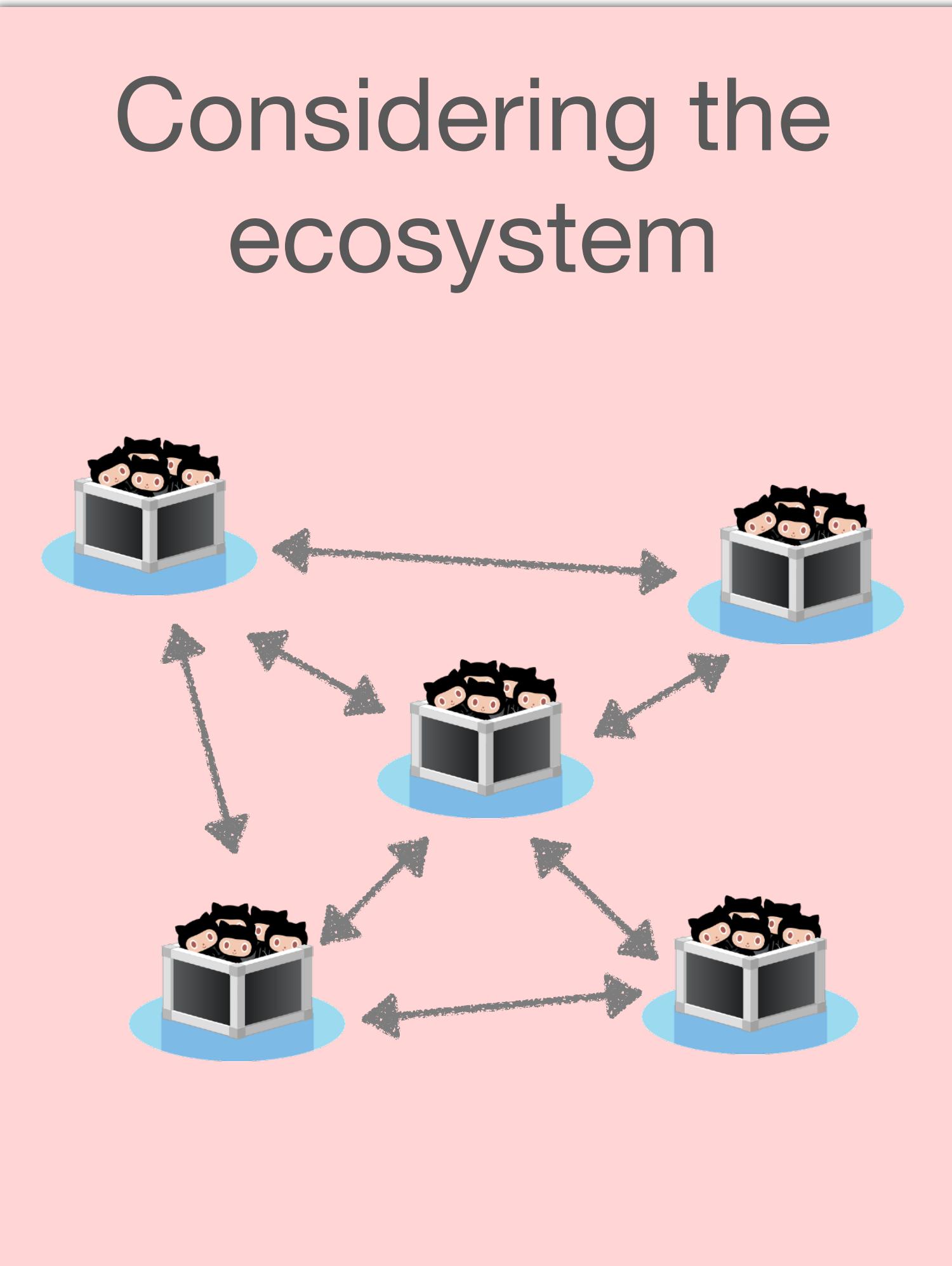
- The Signals that Potential Contributors Look for When Choosing Open-source Projects.
Qiu, S., Li, Yucen., Padala, S., Sarma, A., and Vasilescu, B. *Under review 2019*

Three examples

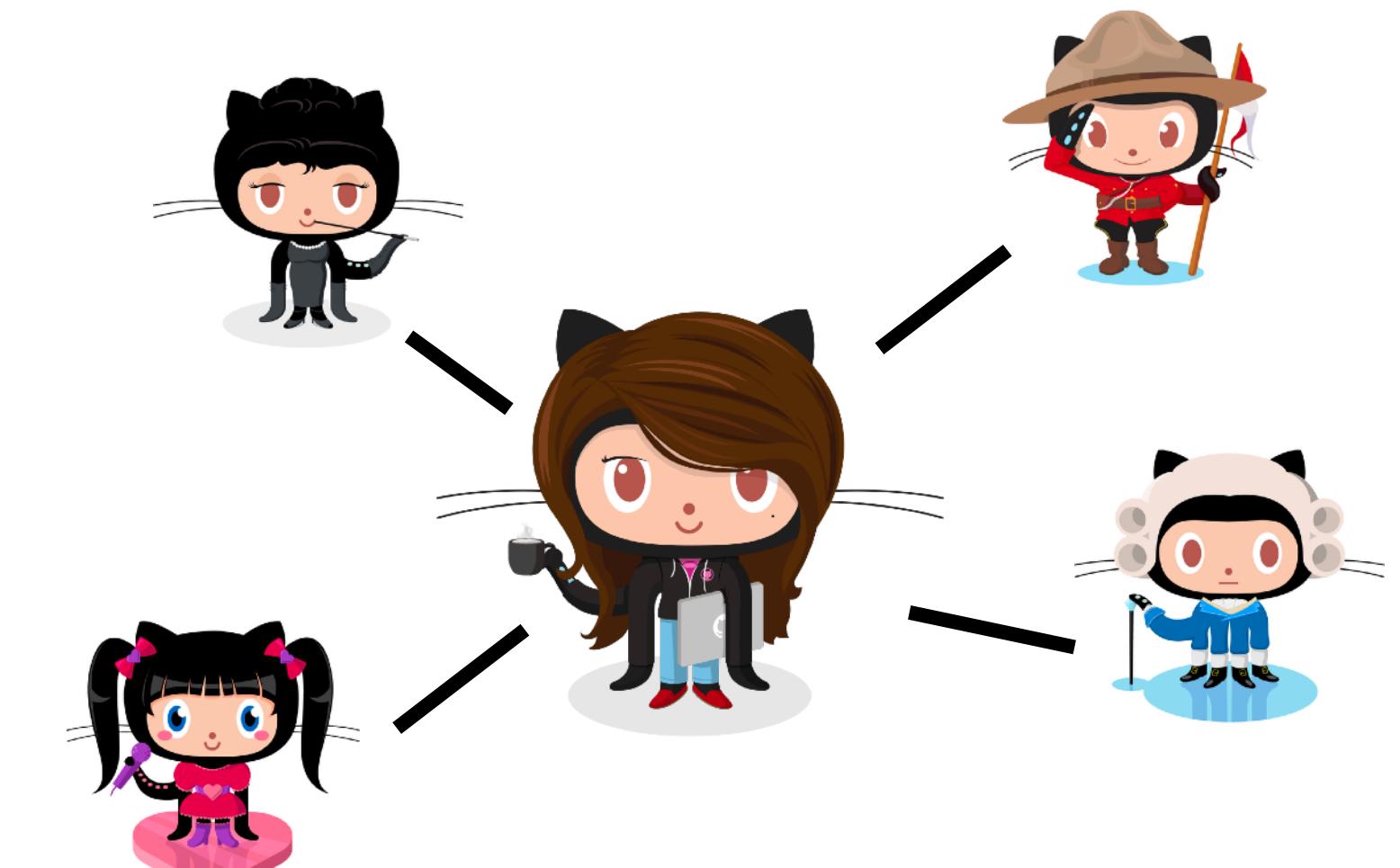
Leveraging transparency



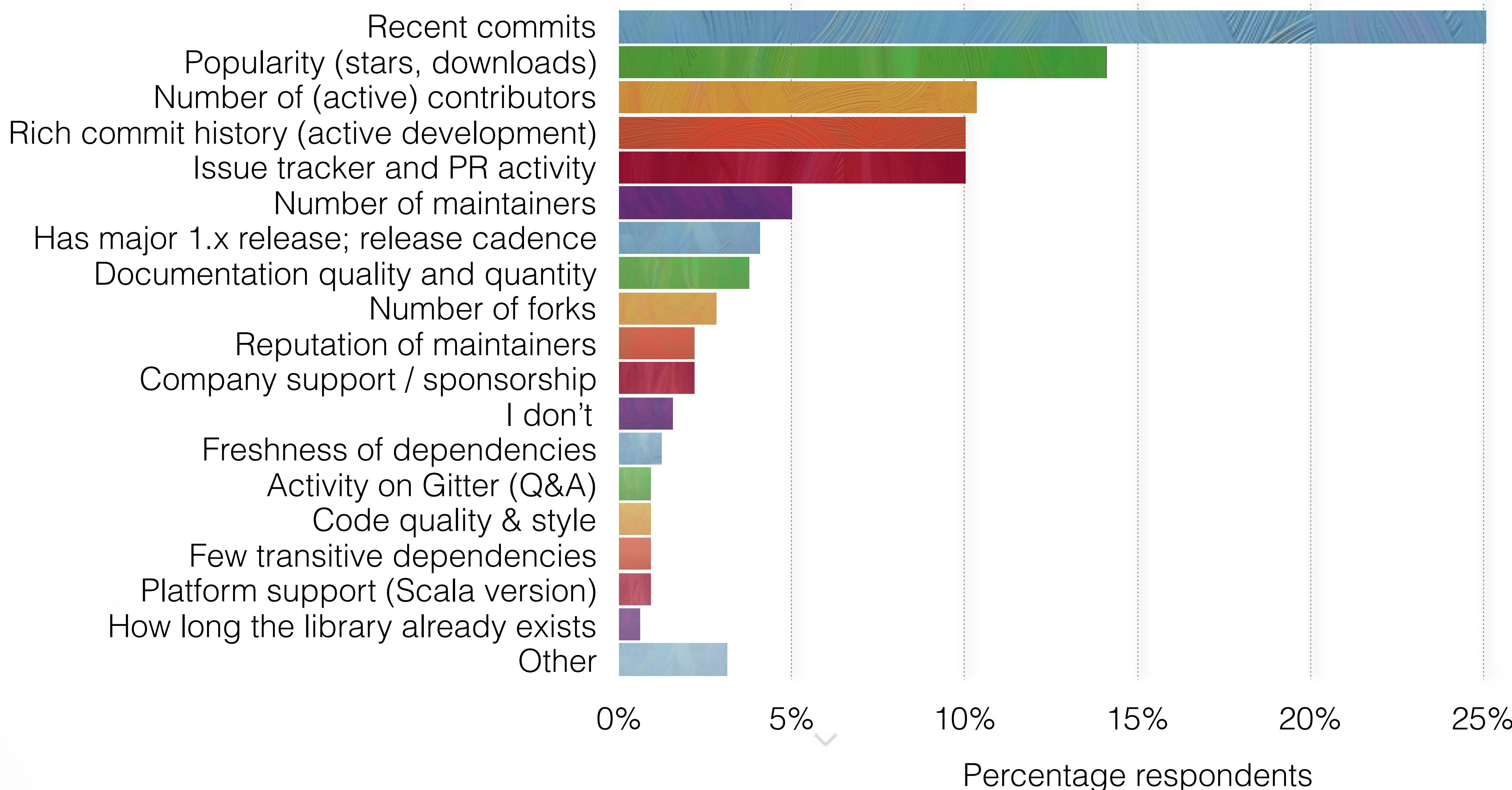
Considering the ecosystem



Building social capital

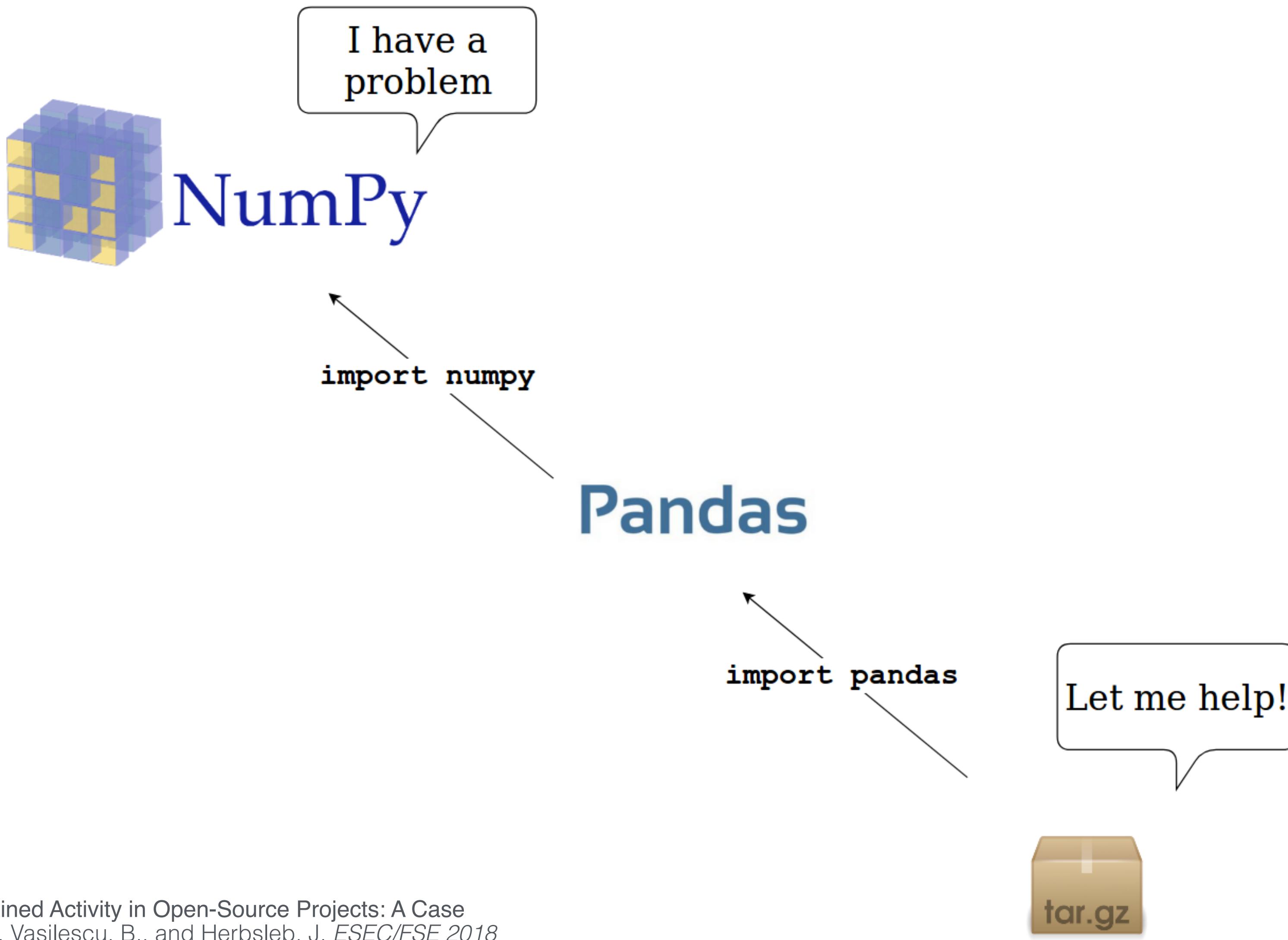
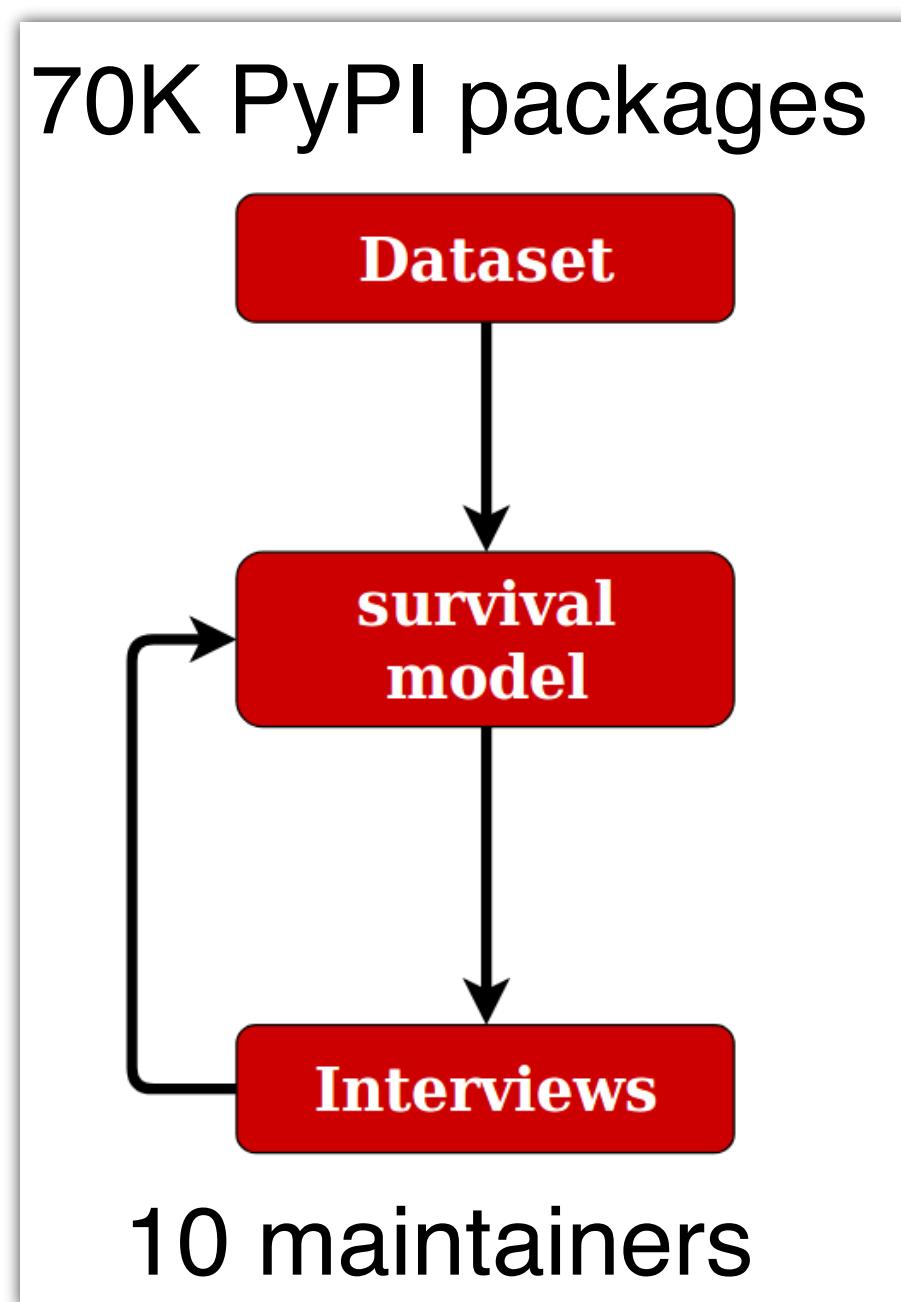


How do you screen open source libraries to make sure they would still be maintained in the future?



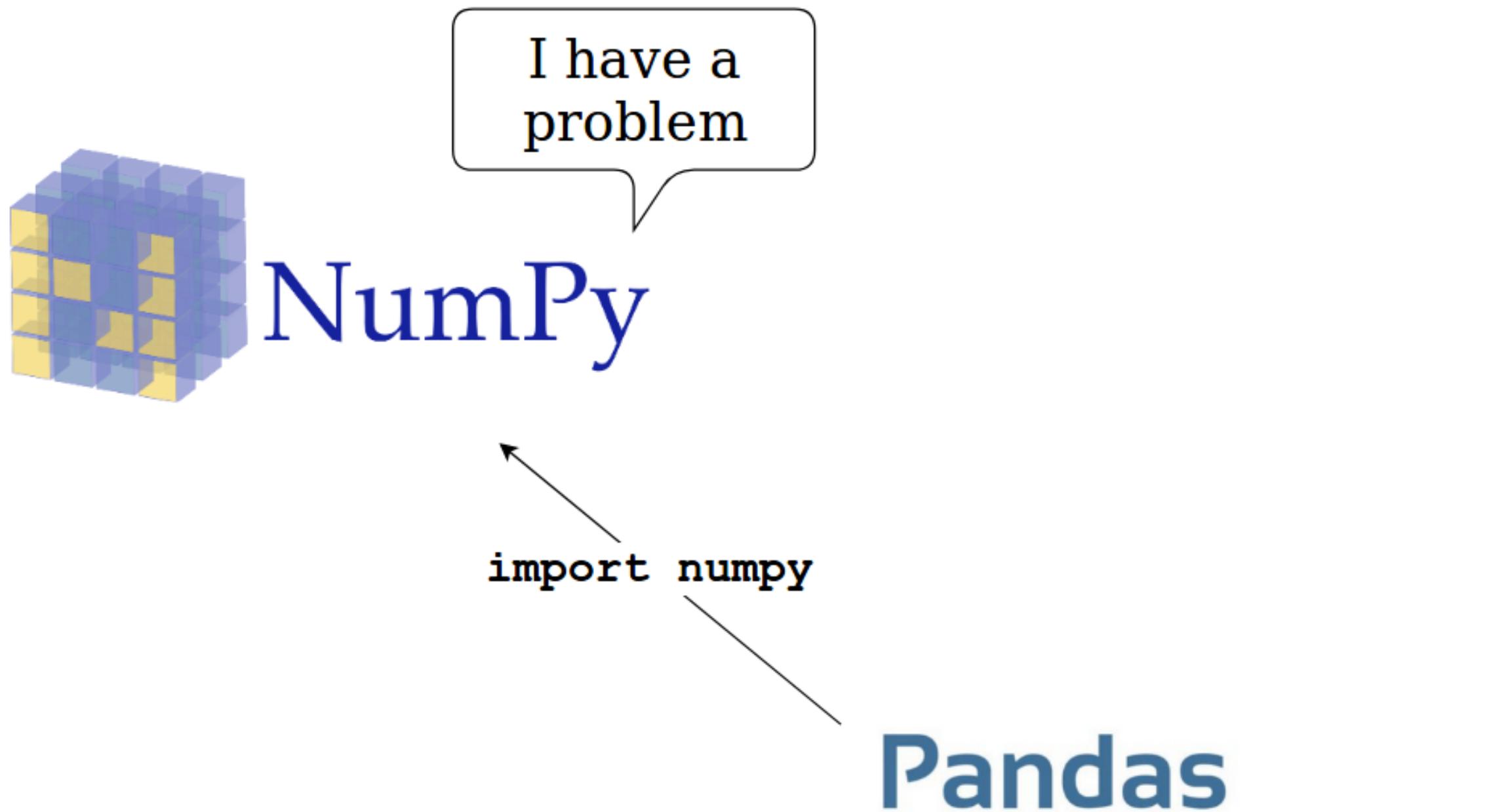
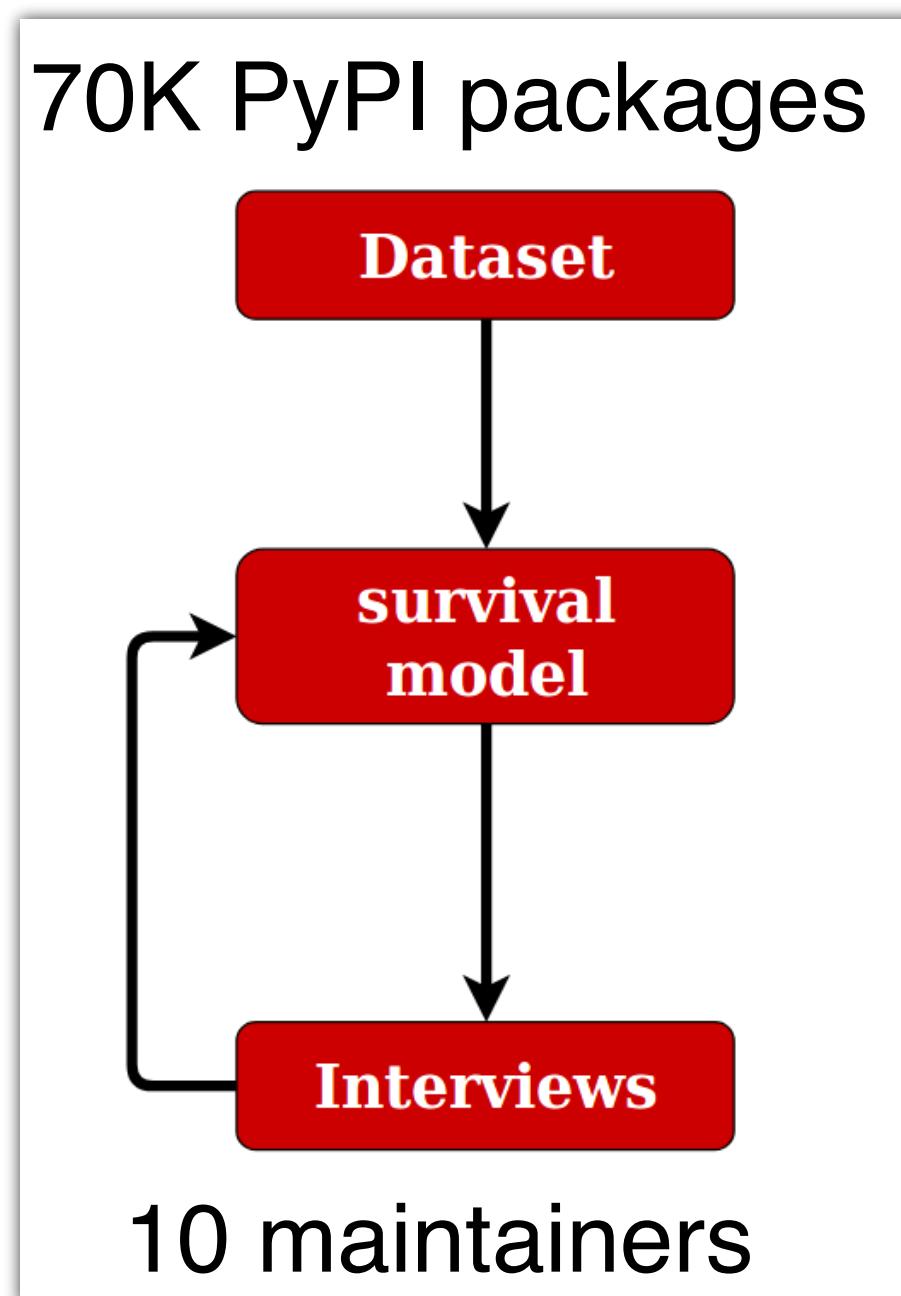
290

Transitive downstream dependencies are



- Ecosystem-Level Determinants of Sustained Activity in Open-Source Projects: A Case Study of the PyPI Ecosystem. Valiev, M., Vasilescu, B., and Herbsleb, J. *ESEC/FSE 2018*

Transitive downstream dependencies are harmful

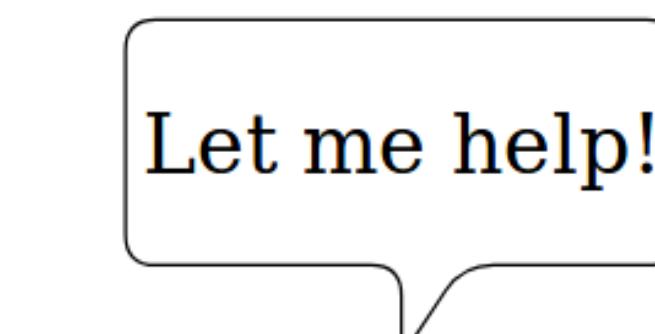


Survival models

Early stage: **-12%** survival
Long term: **-27%** survival

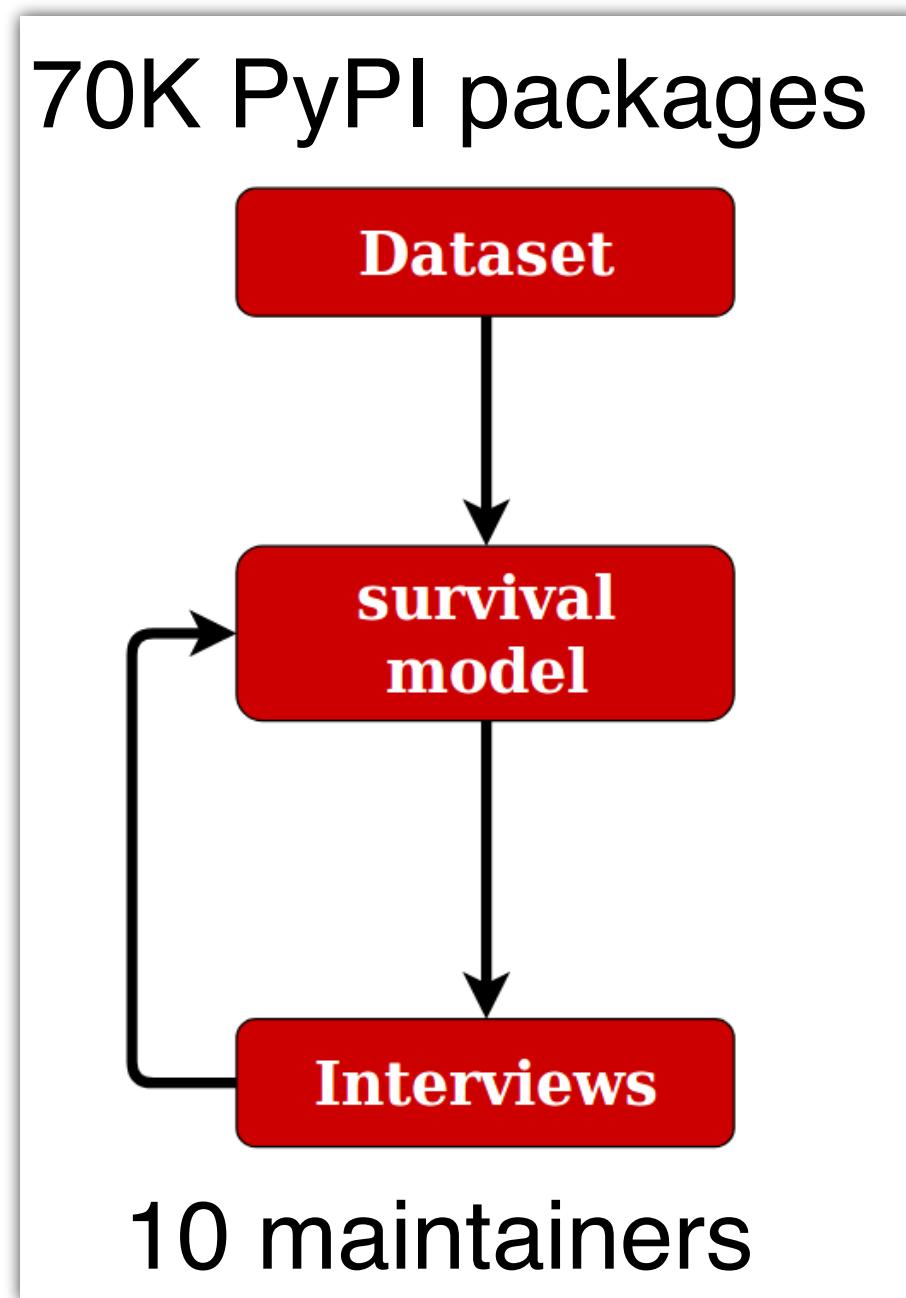
Interviews:

- less likely to fix
- just as likely to complain



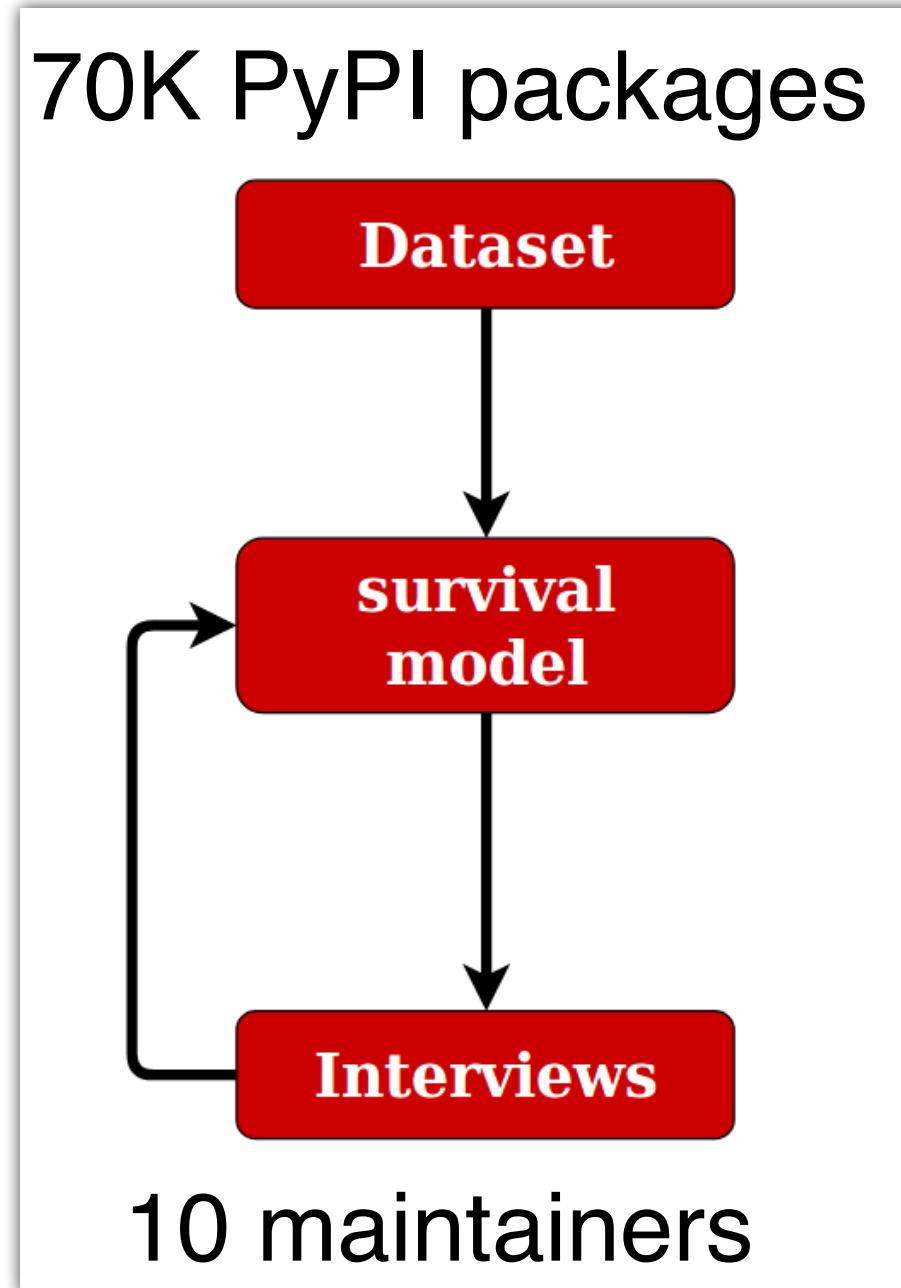
• Ecosystem-Level Determinants of Sustained Activity in Open-Source Projects: A Case Study of the PyPI Ecosystem. Valiev, M., Vasilescu, B., and Herbsleb, J. *ESEC/FSE 2018*

Commercial involvement is



- Ecosystem-Level Determinants of Sustained Activity in Open-Source Projects: A Case Study of the PyPI Ecosystem. Valiev, M., Vasilescu, B., and Herbsleb, J. *ESEC/FSE 2018*

Commercial involvement is harmful



Survival models

Early stage: **-51%** survival

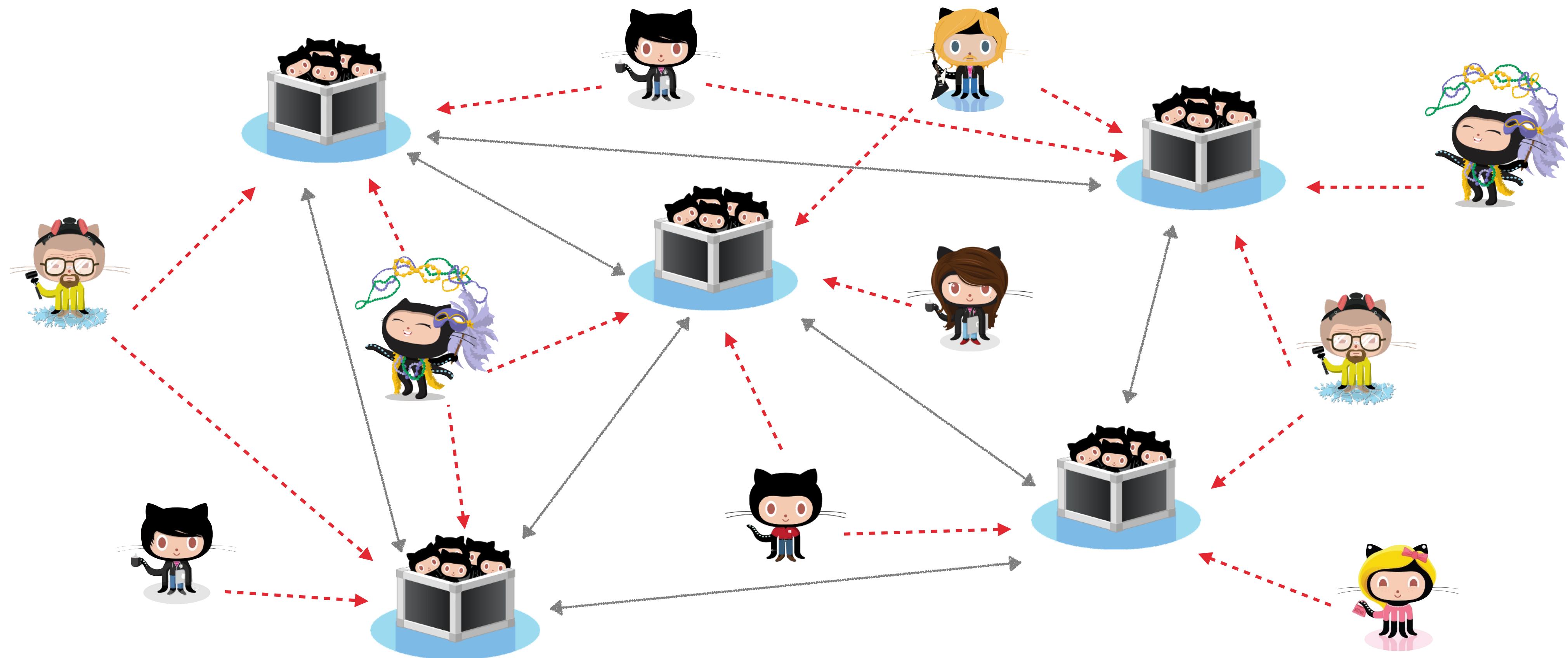
Long term: **-15%** survival

Interviews:

- more resources
- but can withdraw anytime

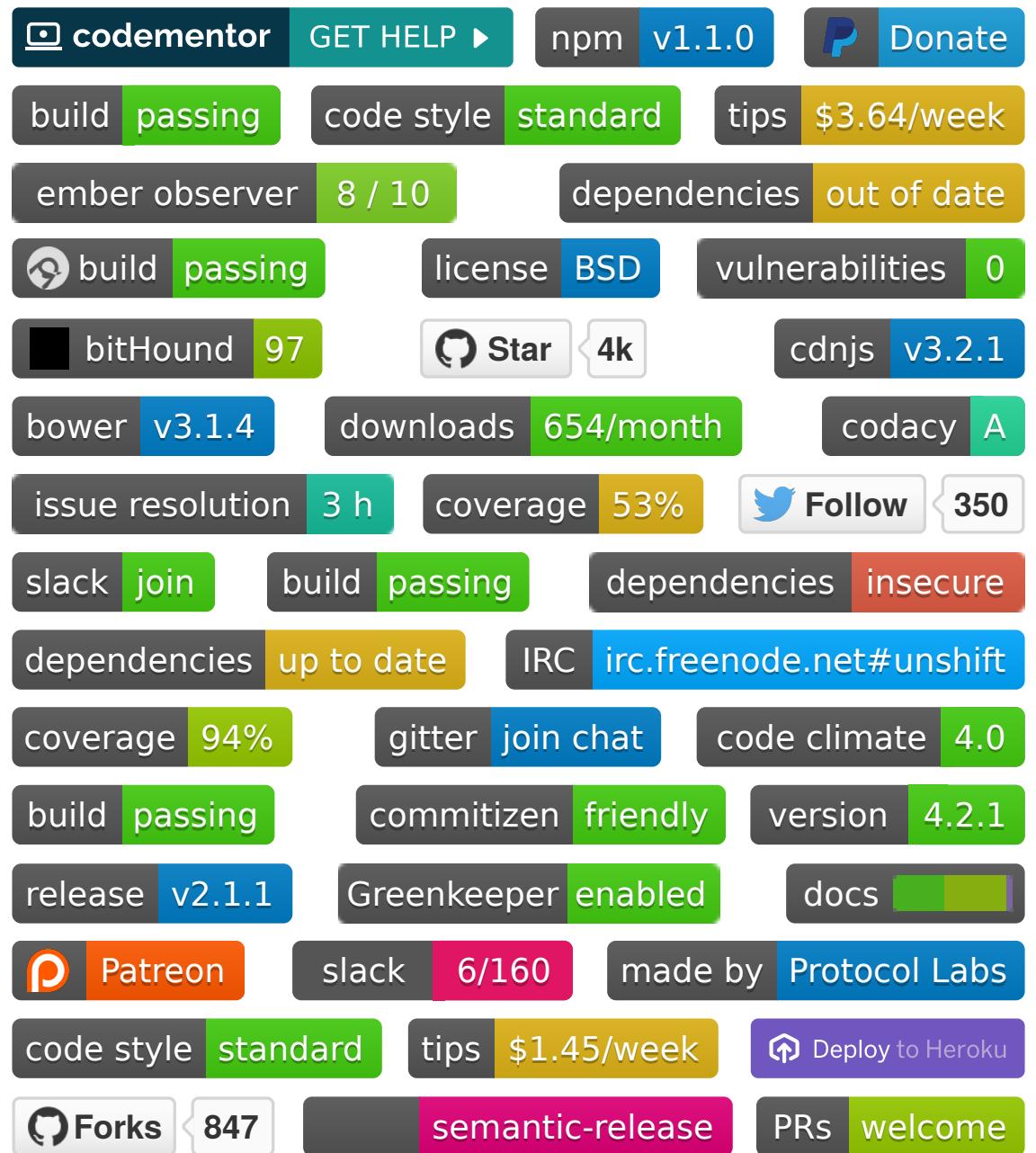
• Ecosystem-Level Determinants of Sustained Activity in Open-Source Projects: A Case Study of the PyPI Ecosystem. Valiev, M., Vasilescu, B., and Herbsleb, J. *ESEC/FSE 2018*

Take away: Ecosystem factors matter too

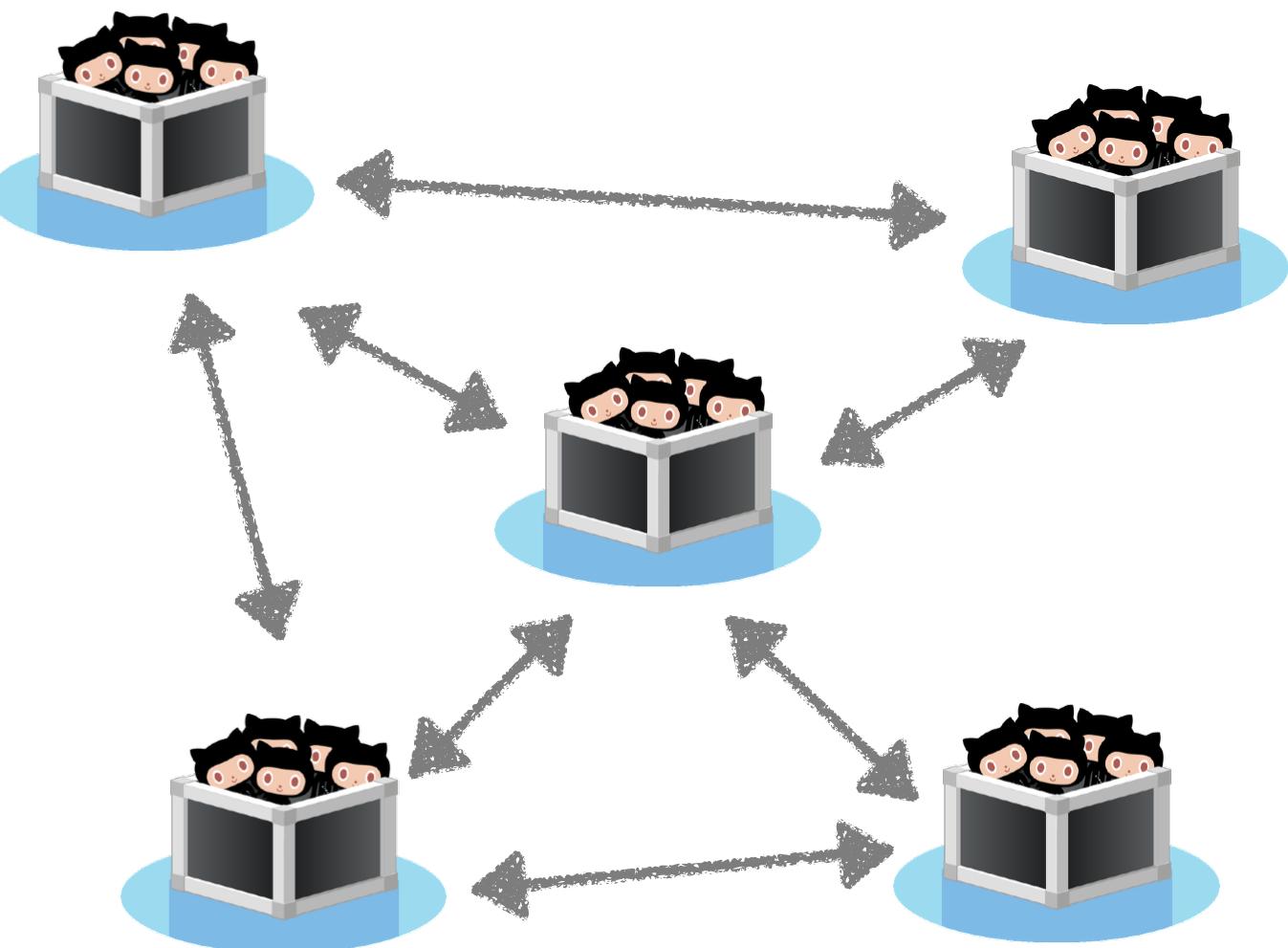


Three examples

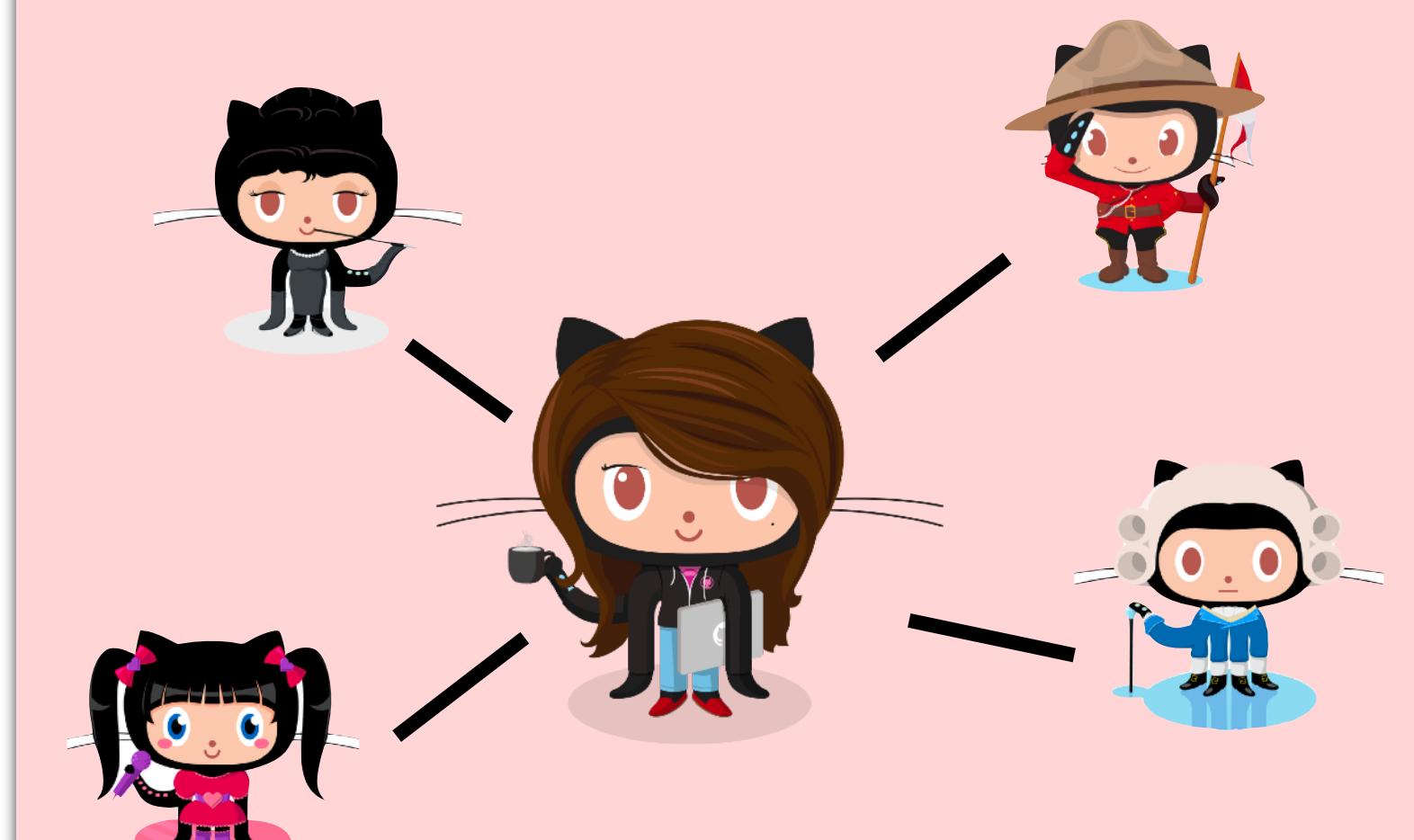
Leveraging transparency



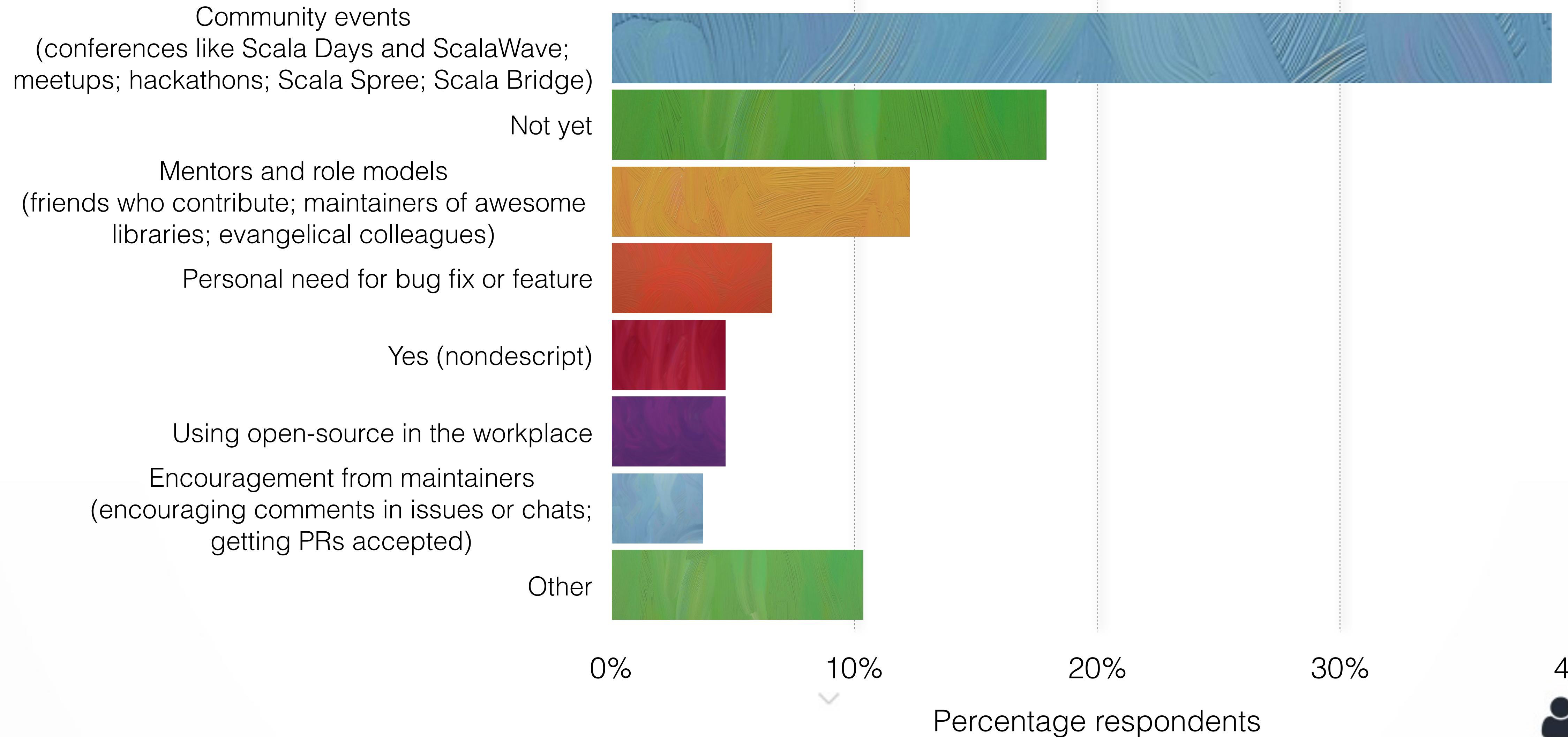
Considering the ecosystem



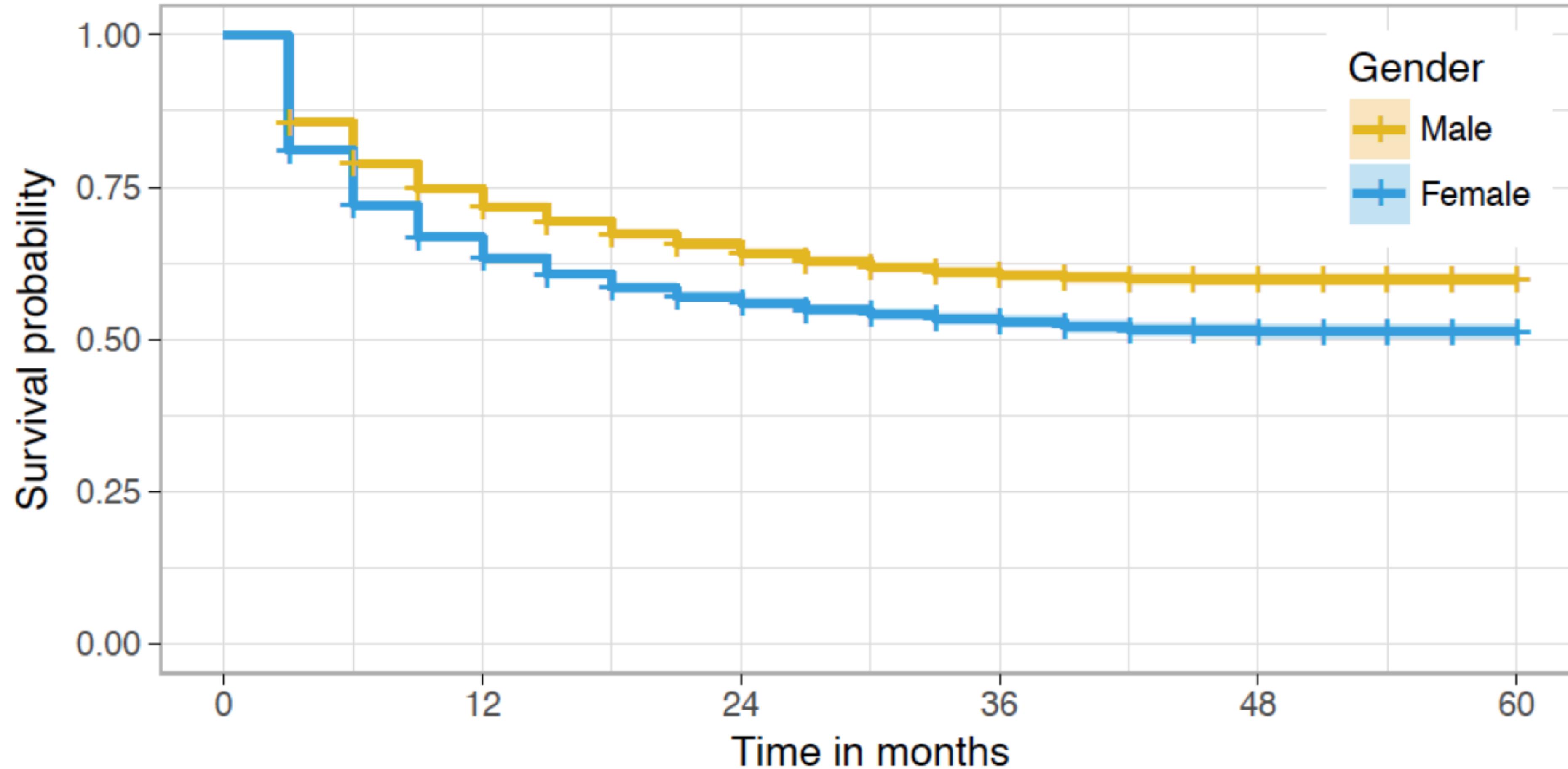
Building social capital



Were there events or people that encouraged you to seriously get involved and stay engaged in open-source?



Women on GitHub disengage earlier than men



- Going Farther Together: The Impact of Social Capital on Sustained Participation in Open Source.
Qiu, H.S., Nolte, A., Brown, A., Serebrenik, A., and Vasilescu, B. *ICSE 2019*

“Sexist behavior in F/LOSS is as constant as it is extreme”

Article



‘Patches don’t have gender’: What is not open in open source software

new media & society
14(4) 669–683
© The Author(s) 2011
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1461444811422887
nms.sagepub.com
SAGE

Dawn Nafus
Intel Labs, USA

Abstract

While open source software development promises a fairer, more democratic model of software production often compared to a gift economy, it also is far more male dominated than other forms of software production. The specific ways F/LOSS instantiates notions of openness in everyday practice exacerbates the exclusion of women. ‘Openness’ is a complex construct that affects more than intellectual property arrangements. It weaves together ideas about authorship, agency, and the circumstances under which knowledge and code can and cannot be exchanged. While open source developers believe technology is orthogonal to the social, notions of openness tie the social to the technical by separating persons from one another and relieving them of obligations that might be created in the course of other forms of gift exchange. In doing so, men monopolize code authorship and simultaneously de-legitimize the kinds of social ties necessary to build mechanisms for women’s inclusion.

“I have used a fake GitHub handle [...] so that people would assume I was male”

Article



new media & society

‘Patches don’t have gender. What is not open in open source software’

Dawn Nafus
Intel Labs, USA

Abstract

While open source software development promises a more democratic form of software production often compared to a gift economy, the reality is more complex than other forms of software production. The specific ways in which the promise of openness in everyday practice exacerbates the exclusions of the open source community construct that affects more than intellectual property rights. It also affects the way ideas about authorship, agency, and the circumstances under which ideas can and cannot be exchanged. While open source development is often seen as tied to the social, notions of openness tie the social to the individual, isolating them from one another and relieving them of obligations that might otherwise be associated with forms of gift exchange. In doing so, men monopolize the kinds of social ties necessary to build and maintain the open source community.

Perceptions of Diversity on GitHub: A User Survey

Bogdan Vasilescu
University of California, Davis
vasilescu@ucdavis.edu

Vladimir Filkov
University of California, Davis
filkov@cs.ucdavis.edu

Alexander Serebrenik
Eindhoven University of Technology
a.serebrenik@tue.nl

Abstract—Understanding one’s work environment is important for one’s success, especially when working in teams. In virtual collaborative environments this amounts to being aware of the technical and social attributes of one’s team members. Focusing on Open Source Software teams, naturally very diverse both socially and technically, we report the results of a user survey that tries to resolve how teamwork and individual attributes are perceived by developers collaborating on GITHUB, and how those perceptions influence their work. Our findings can be used as complementary data to quantitative studies of developers’ behavior on GITHUB.

I. INTRODUCTION

Software development is technical and knowledge-intensive, but also human-centric and collaborative, benefiting from the social attributes of the people involved. Open Source Software (OSS) communities, in particular, tend to be quite diverse, with contributors ranging from professional developers to volunteers, all with varied personalities, educational and cultural backgrounds, age, gender, and expertise. Yet, despite participating in a very decentralized process, and despite this diversity, OSS teams often succeed to work together effectively and productively [1], [2].

attributes (e.g., gender, tenure, political views) on the overall work environment. Our previous study [7] was, to the best of our knowledge, the first to consider effects of gender diversity on productivity and turnover in OSS communities, and one of the very few studies of diversity in general in OSS or other online peer production systems (e.g., [14]–[16]).

In this paper we offer a qualitative perspective of diversity in software teams: we report the results of a user survey that tries to resolve how teamwork and individual attributes are perceived by developers collaborating on GITHUB, and how those perceptions influence their work. We address a number of research questions, as discussed next.

OSS teams are typically more fluid and less tangible than their offline counterparts. They tend to form and dissolve organically around the task at hand, facing high turnover [17], while interactions between members are often limited to online channels [18]. In addition, GITHUB’s implementation of the pull-based development model [19] enables anyone to submit changes to any repository with minimal effort, through pull requests (the so-called “drive-by” commits [13]). We wish to understand whether this unprecedented low barrier to entry for

“I have used a fake GitHub handle [...] so that people would assume I was male”

Article

A small square graphic showing a spiral pattern of binary digits (0s and 1s) on a dark background.

new media & society

‘Patches don’t have gender What is not open in open source software’

Dawn Nafus
Intel Labs, USA

Abstract

While open source software development promises a more democratic form of software production often compared to a gift economy than other forms of software production. The specific nature of openness in everyday practice exacerbates the exclusionary construct that affects more than intellectual property ideas about authorship, agency, and the circumstances in which ideas can and cannot be exchanged. While open source developers are tied to the social, notions of openness tie the social to the individual, one another and relieving them of obligations that may limit forms of gift exchange. In doing so, men monopolize the space and de-legitimize the kinds of social ties necessary to build a sustainable community.

Perceptions of Diversity on GitHub

Bogdan Vasilescu
University of California, Davis
vasilescu@ucdavis.edu

Vladimir Filkov
University of California, Davis
filkov@cs.ucdavis.edu

Abstract—Understanding one’s work environment is important for one’s success, especially when working in teams. In virtual collaborative environments this amounts to being aware of the technical and social attributes of one’s team members. Focusing on Open Source Software teams, naturally very diverse both socially and technically, we report the results of a user survey that tries to resolve how teamwork and individual attributes are perceived by developers collaborating on GitHub, and how those perceptions influence their work. Our findings can be used as complementary data to quantitative studies of developers’ behavior on GitHub.

I. INTRODUCTION

Software development is technical and knowledge-intensive, but also human-centric and collaborative, benefiting from the social attributes of the people involved. Open Source Software (OSS) communities, in particular, tend to be quite diverse, with contributors ranging from professional developers to volunteers, all with varied personalities, educational and cultural backgrounds, age, gender, and expertise. Yet, despite participating in a very decentralized process, and despite this diversity, OSS teams often succeed to work together effectively and productively [1], [2].

attribute of work environment our knowledge on products the very online platform.

In this paper, in software tries to perceive those people of research.

OSS teams their off-line organizations while interacting channels [10].

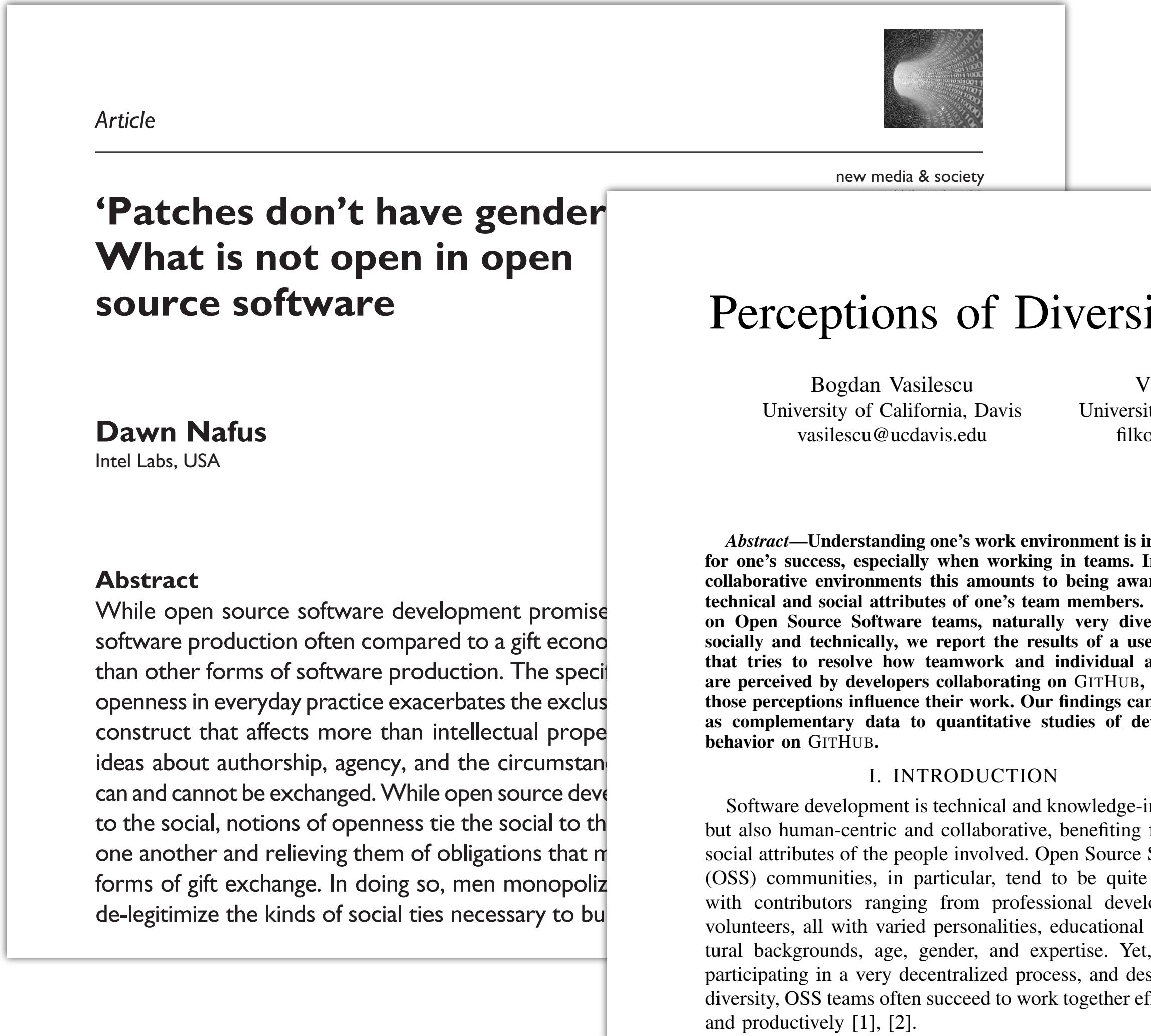
In addition, GitHub’s implementation of the pull-based development model [19] enables anyone to submit changes to any repository with minimal effort, through pull requests (the so-called “drive-by” commits [13]). We wish to understand whether this unprecedented low barrier to entry for

Developers are aware of each other’s gender

Which of the following characteristics of your team members are you aware of?

- 74% • Programming skills
- 48% • Gender
- 45% • Real name
- 42% • Social skills
- 40% • Country of residence
- 39% • Personality
- 31% • Reputation as programmer
- 30% • Ethnicity
- 30% • Employment
- 28% • GitHub experience
- 26% • Educational level
- 23% • Age
- 11% • Hobbies
- 4% • Political views

Pull request acceptance rates are lower when gender is apparent



Article

'Patches don't have gender. What is not open in open source software'

Dawn Nafus
Intel Labs, USA

Abstract
While open source software development promises a more collaborative and meritocratic way of producing software, it has been argued that the social dynamics of open source software production often compare to a gift economy rather than other forms of software production. The specific nature of openness in everyday practice exacerbates the exclusions that construct a social hierarchy that affects more than intellectual property rights. Ideas about authorship, agency, and the circumstances under which ideas can and cannot be exchanged. While open source development is often seen as being more open to the social, notions of openness tie the social to the technical and the individual to one another and relieving them of obligations that may be associated with other forms of gift exchange. In doing so, men monopolize the kinds of social ties necessary to build OSS teams.

Perceptions of Diversity

Bogdan Vasilescu
University of California, Davis
vasilescu@ucdavis.edu

Vlad Filkov
University of California, Davis

Abstract—Understanding one's work environment is important for one's success, especially when working in teams. In collaborative environments this amounts to being aware of the technical and social attributes of one's team members. Focusing on Open Source Software teams, naturally very diverse both socially and technically, we report the results of a user study that tries to resolve how teamwork and individual attributes are perceived by developers collaborating on GITHUB, and how those perceptions influence their work. Our findings can be used as complementary data to quantitative studies of developer behavior on GITHUB.

I. INTRODUCTION

Software development is technical and knowledge-intensive, but also human-centric and collaborative, benefiting from the social attributes of the people involved. Open Source Software (OSS) communities, in particular, tend to be quite diverse, with contributors ranging from professional developers to volunteers, all with varied personalities, educational and cultural backgrounds, age, gender, and expertise. Yet, despite participating in a very decentralized process, and despite this diversity, OSS teams often succeed to work together effectively and productively [1], [2].



Gender differences and bias in open source: pull request acceptance of women versus men

Josh Terrell¹, Andrew Kofink², Justin Middleton², Clarissa Rainear², Emerson Murphy-Hill², Chris Parnin² and Jon Stallings³

¹ Department of Computer Science, California Polytechnic State University—San Luis Obispo, San Luis Obispo, CA, United States

² Department of Computer Science, North Carolina State University, Raleigh, NC, United States

³ Department of Statistics, North Carolina State University, Raleigh, NC, United States

ABSTRACT

Biases against women in the workplace have been documented in a variety of studies. This paper presents a large scale study on gender bias, where we compare acceptance rates of contributions from men versus women in an open source software community. Surprisingly, our results show that women's contributions tend to be accepted more often than men's. However, for contributors who are outsiders to a project and their gender is identifiable, men's acceptance rates are higher. Our results suggest that although women on GitHub may be more competent overall, bias against them exists nonetheless.

OSS teams are typically more fluid and less tangible than their offline counterparts. They tend to form and dissolve organically around the task at hand, facing high turnover [17], while interactions between members are often limited to online channels [18]. In addition, GITHUB's implementation of the pull-based development model [19] enables anyone to submit changes to any repository with minimal effort, through pull requests (the so-called “drive-by” commits [13]). We wish to understand whether this unprecedented low barrier to entry for

Wrong incentives? “Longest streak” backlash

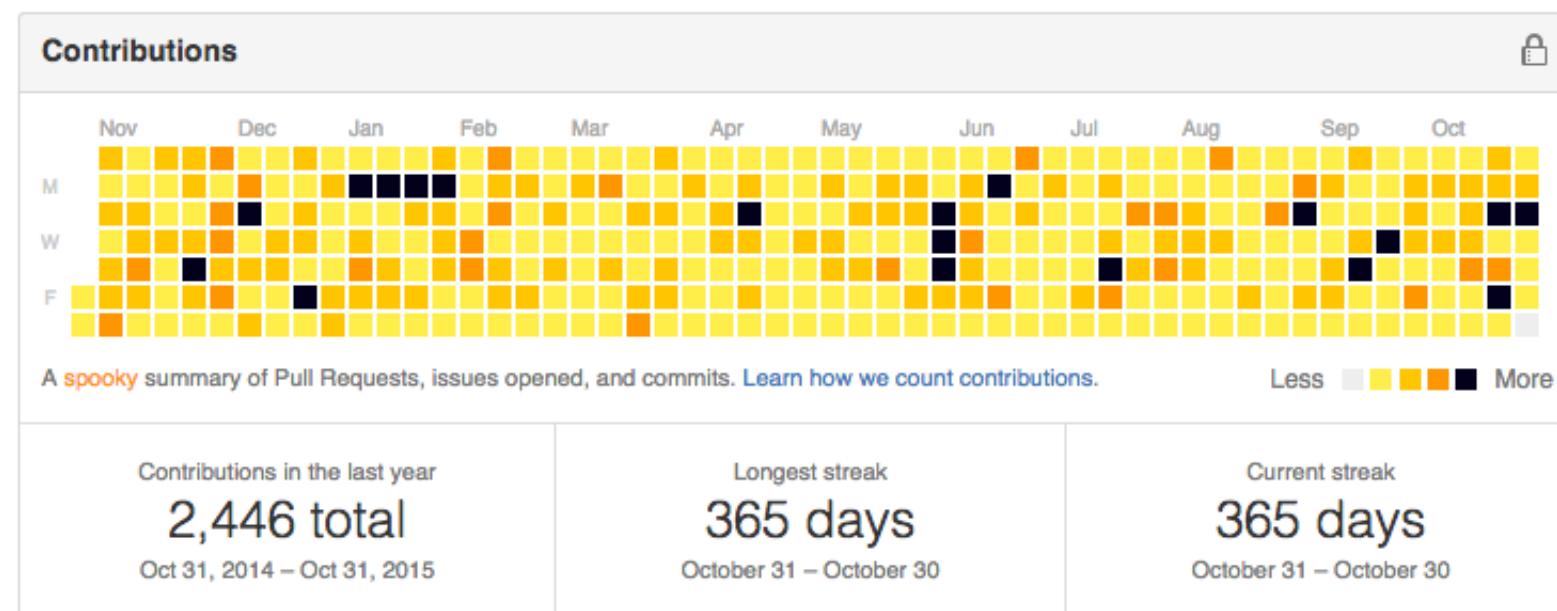
365 days streak on GitHub



Harry Ng [Follow](#)

Oct 31, 2015 · 2 min read

On the day while I am going to celebrate my continuous contribution to GitHub for 365 days, I suddenly found out the colour of the graph changes from green to yellow-orange in colour.



It was a plan started early last year, when I saw a HackerNews about [the longest streak on GitHub \(500 days\)](#). I am so impressed by that, and started to make some achievements by myself. I then started the practice in around June.

<https://medium.com/@harryworld/365-days-streak-on-github-4ceb588ba4ba>

Contribution graph can be harmful to contributors #627

[! Open](#)

mxsasha opened this issue on Apr 1, 2016 · 189 comments



mxsasha commented on Apr 1, 2016

A common well-being issue in open-source communities is the tendency of people to over-commit. Many contributors care deeply, at the risk of saying yes too often harming their well-being. Open-source communities are especially at risk, because many contributors work next to a full-time job.

...

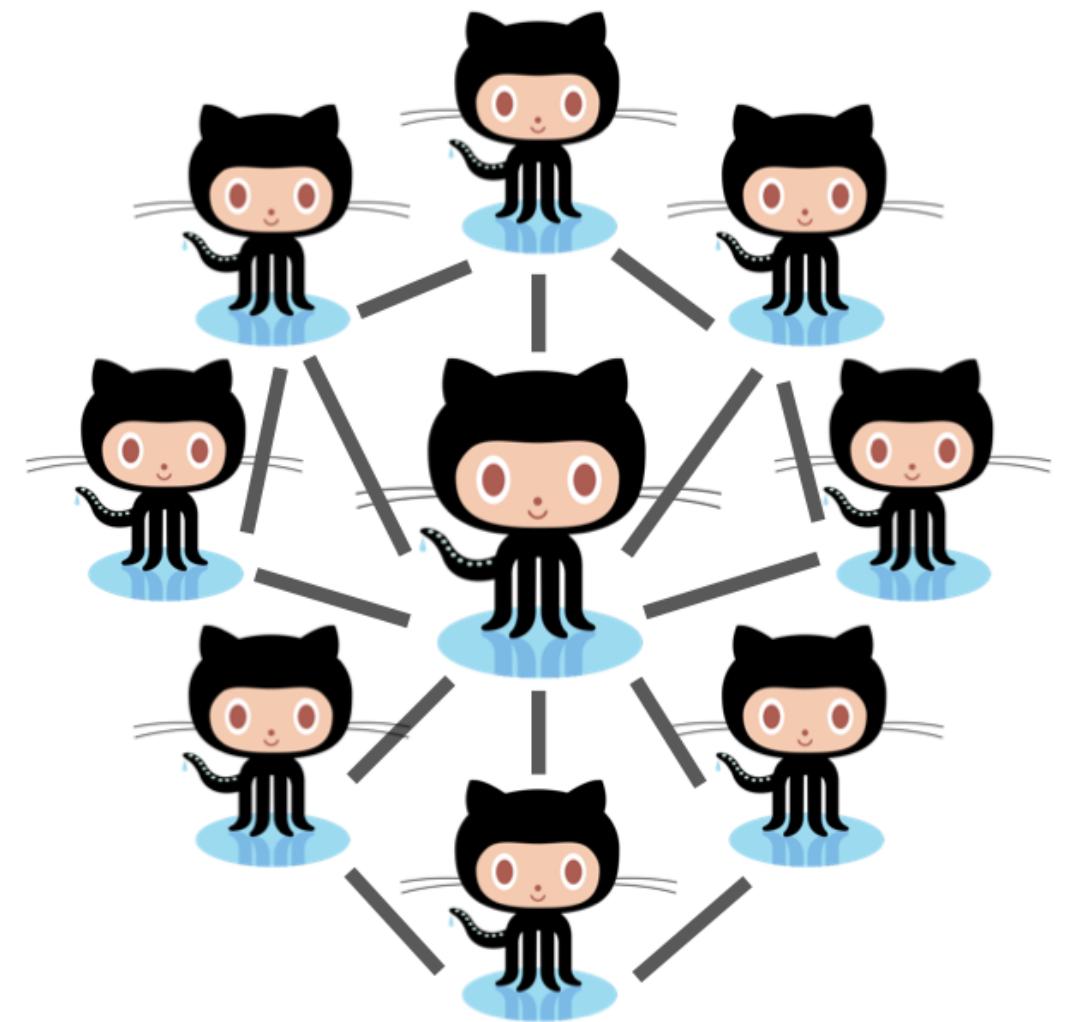
Any mechanism in our community that motivates people to avoid taking breaks and avoid stepping back, can be harmful to the well-being of contributors and is thereby harmful to open source as a whole. Even though it was probably introduced with the best intentions. If our interests are really in supporting open-source long-term, this graph should be removed or substantially changed so that it no longer punishes healthy behaviour. For example, what if we would give people achievements for taking breaks instead of working non-stop?

I therefore want to ask you to consider removing or substantially changing the contribution graph and its related statistics, to help guard the well-being of the contributors and the communities.

I also wrote about this in a bit more detail on my blog: <http://erik.io/blog/2016/04/01/how-github-contribution-graph-is-harmful/>

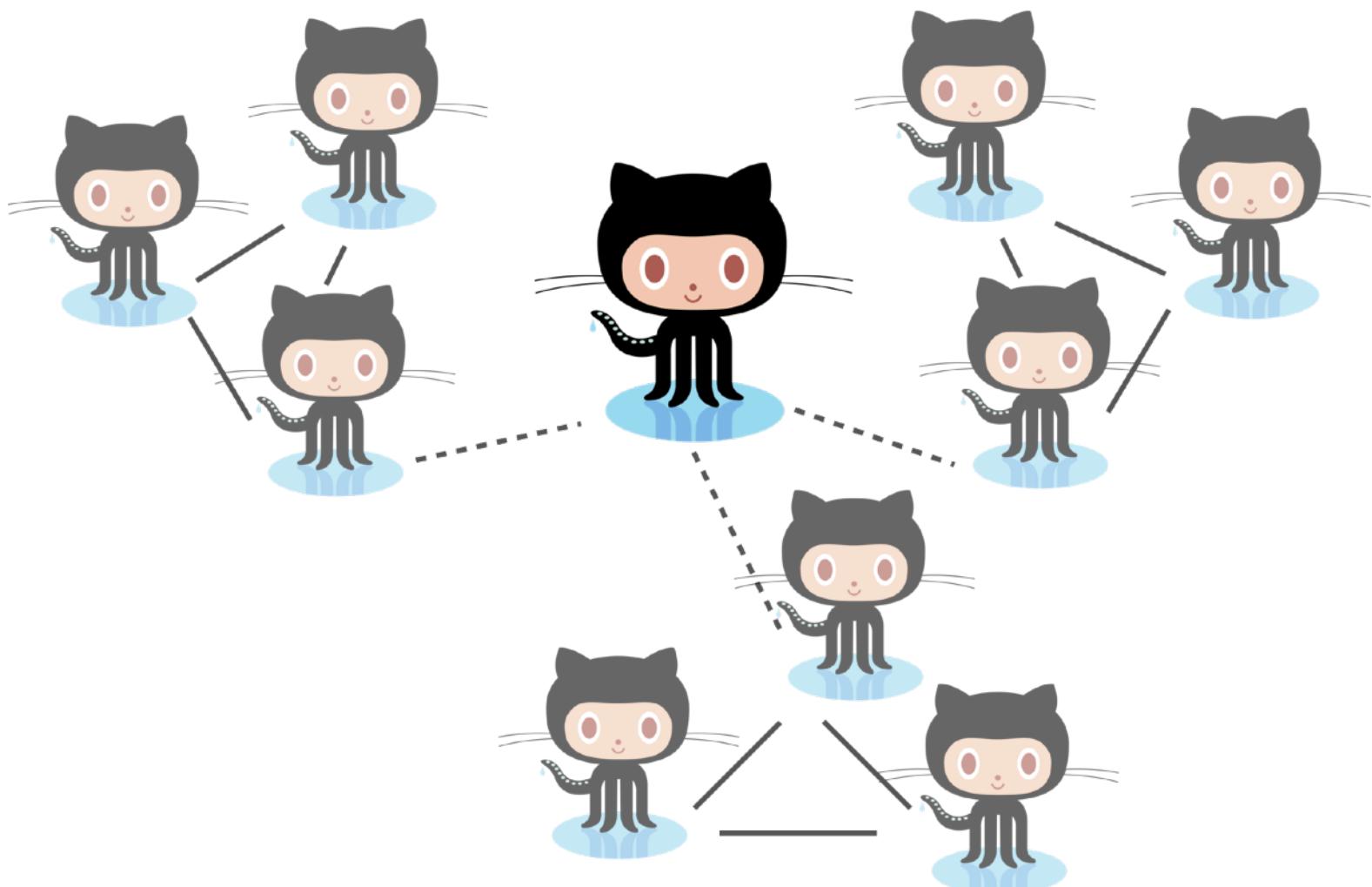
Social capital theory explains long-term engagement

Bonding social capital:
benefiting from strongly
connected network



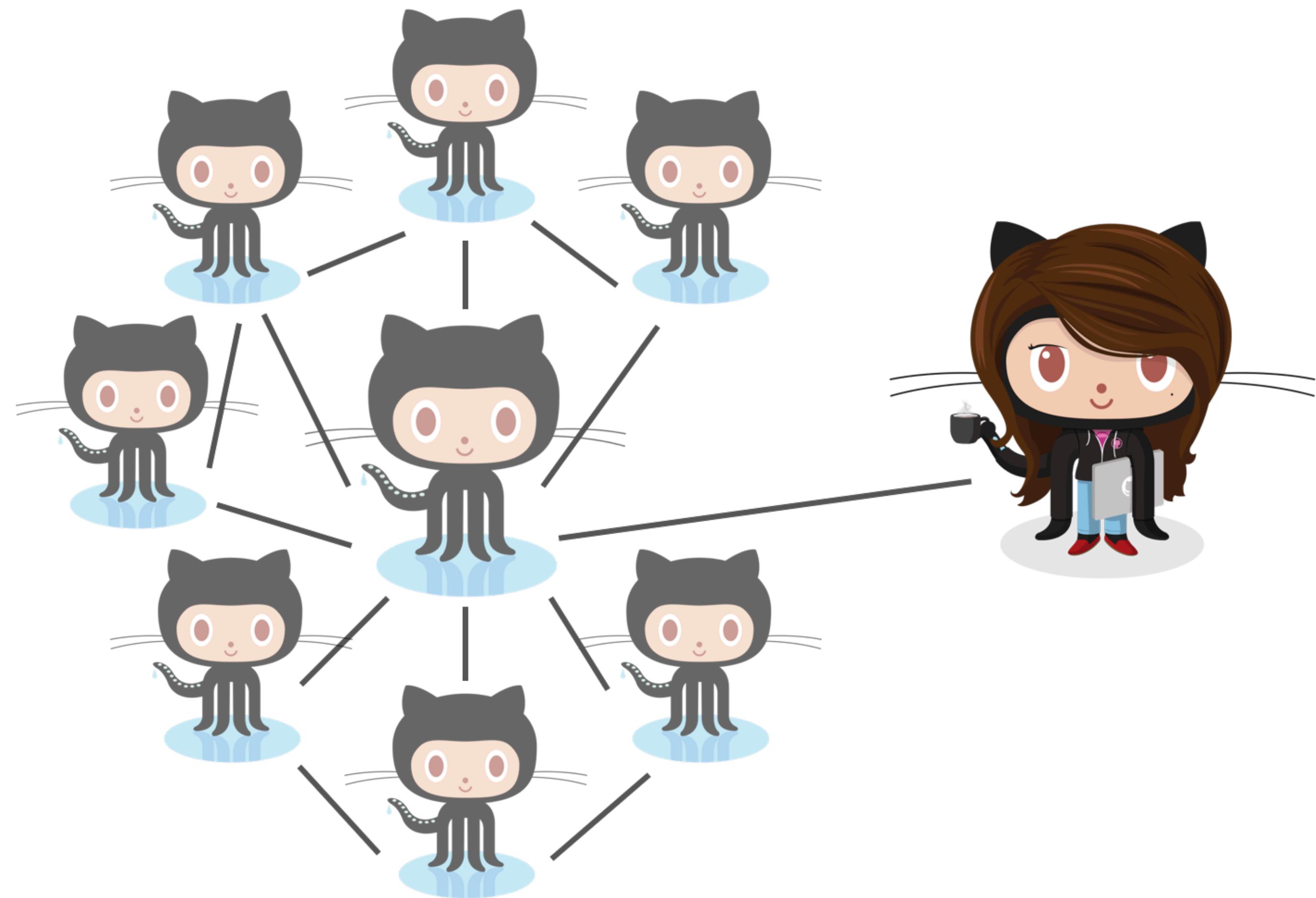
Willingness to continue
(Coleman, 1990)

Bridging social capital:
benefiting from network with
diverse info



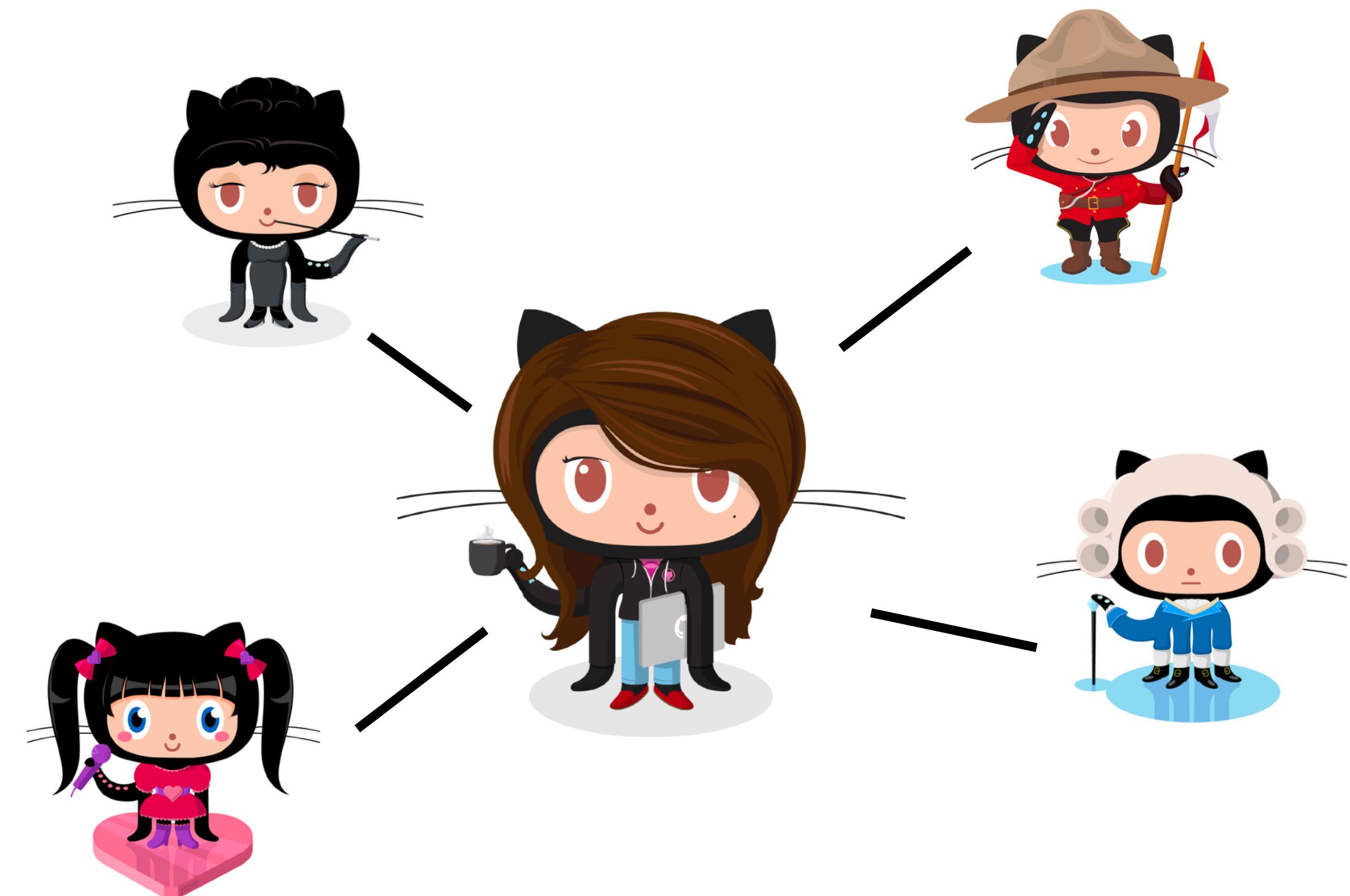
Opportunity to continue
(Burt, 1998, 2001)

Cohesive networks might foster discrimination / exclusion

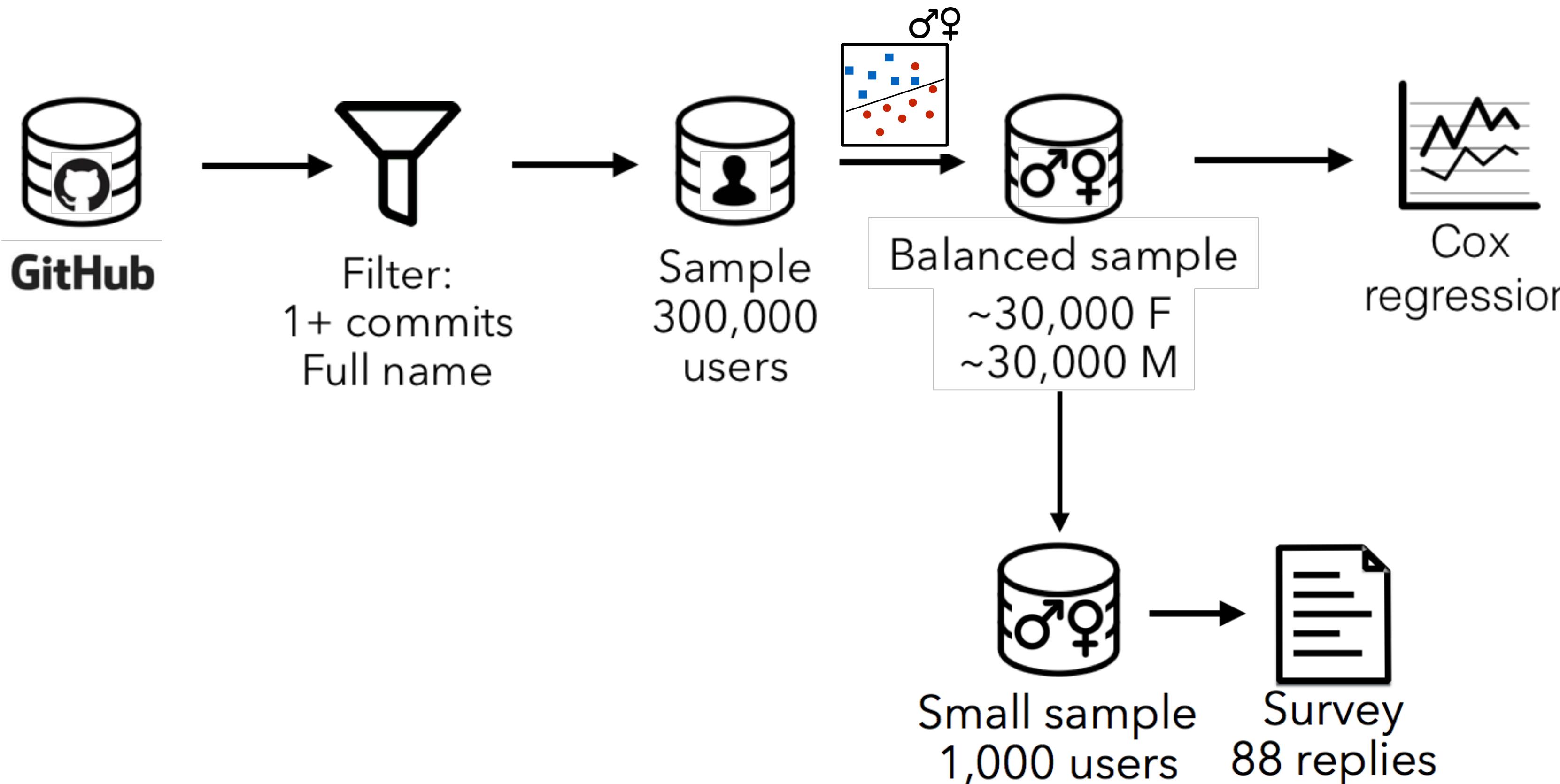


Being part of teams with more diverse information ~ more prolonged engagement, esp. for women

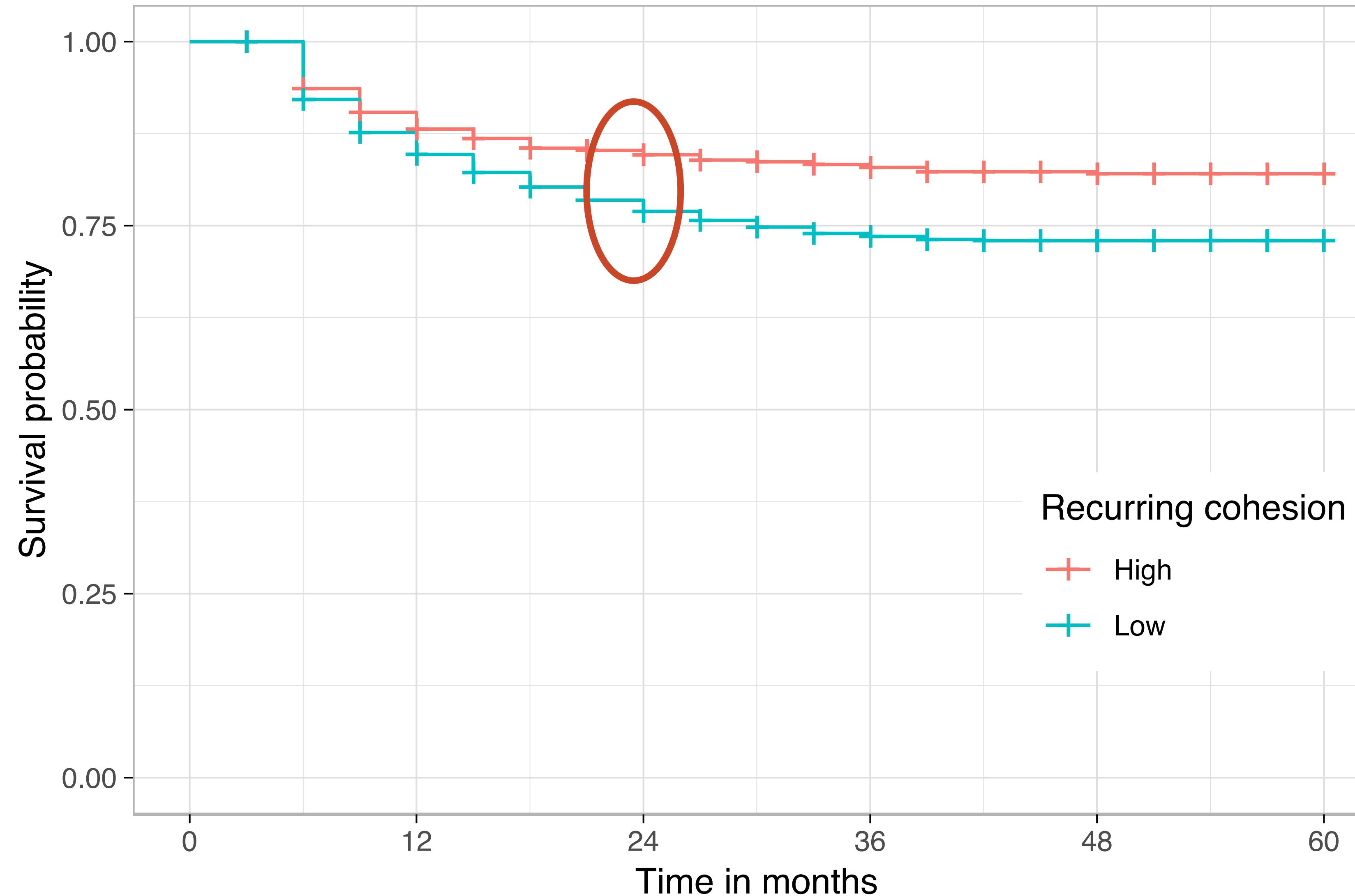
Information diversity should
reduce the risk of demographic-
based echo chambers.



Large-scale mixed-methods study

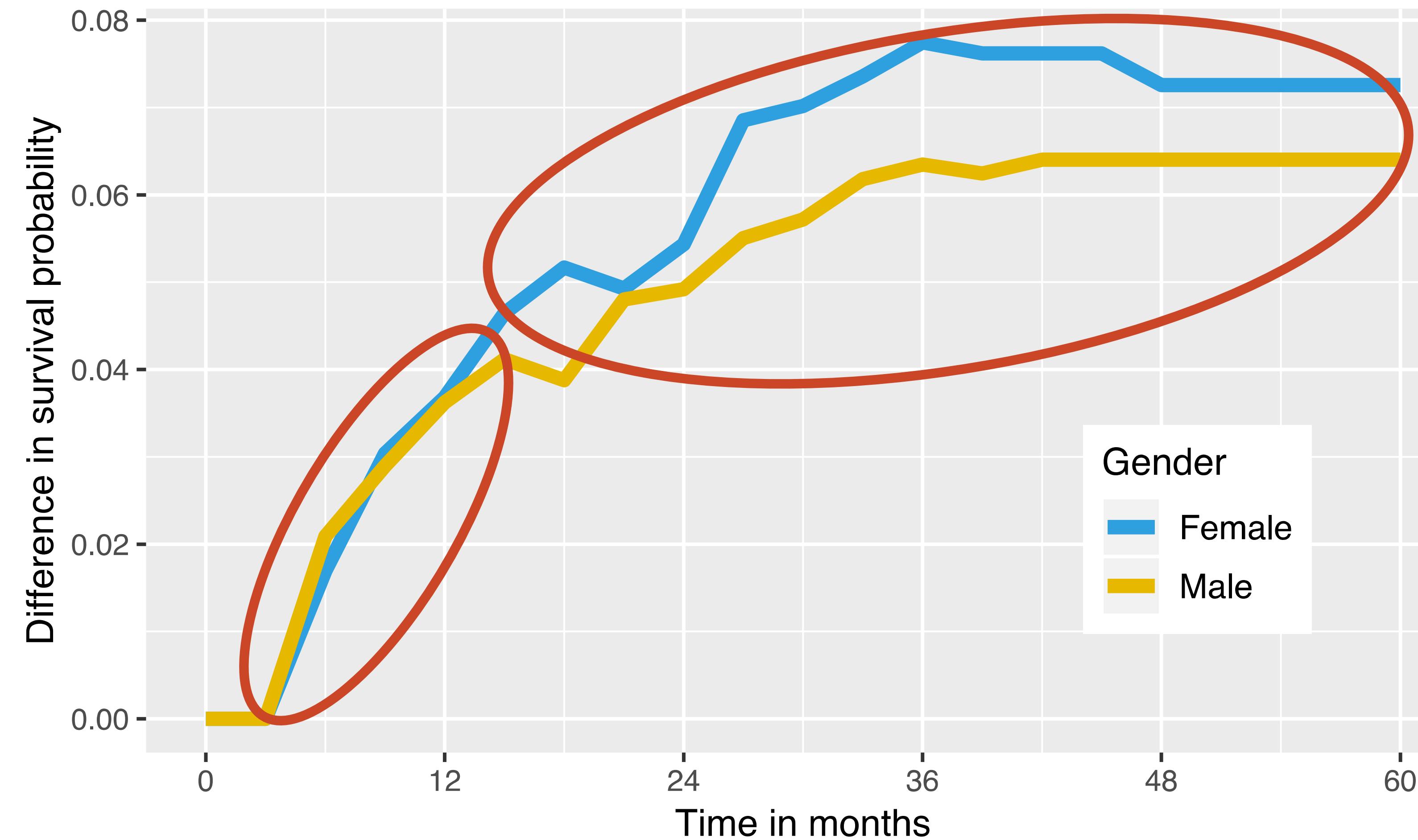


More social capital ~ more prolonged engagement

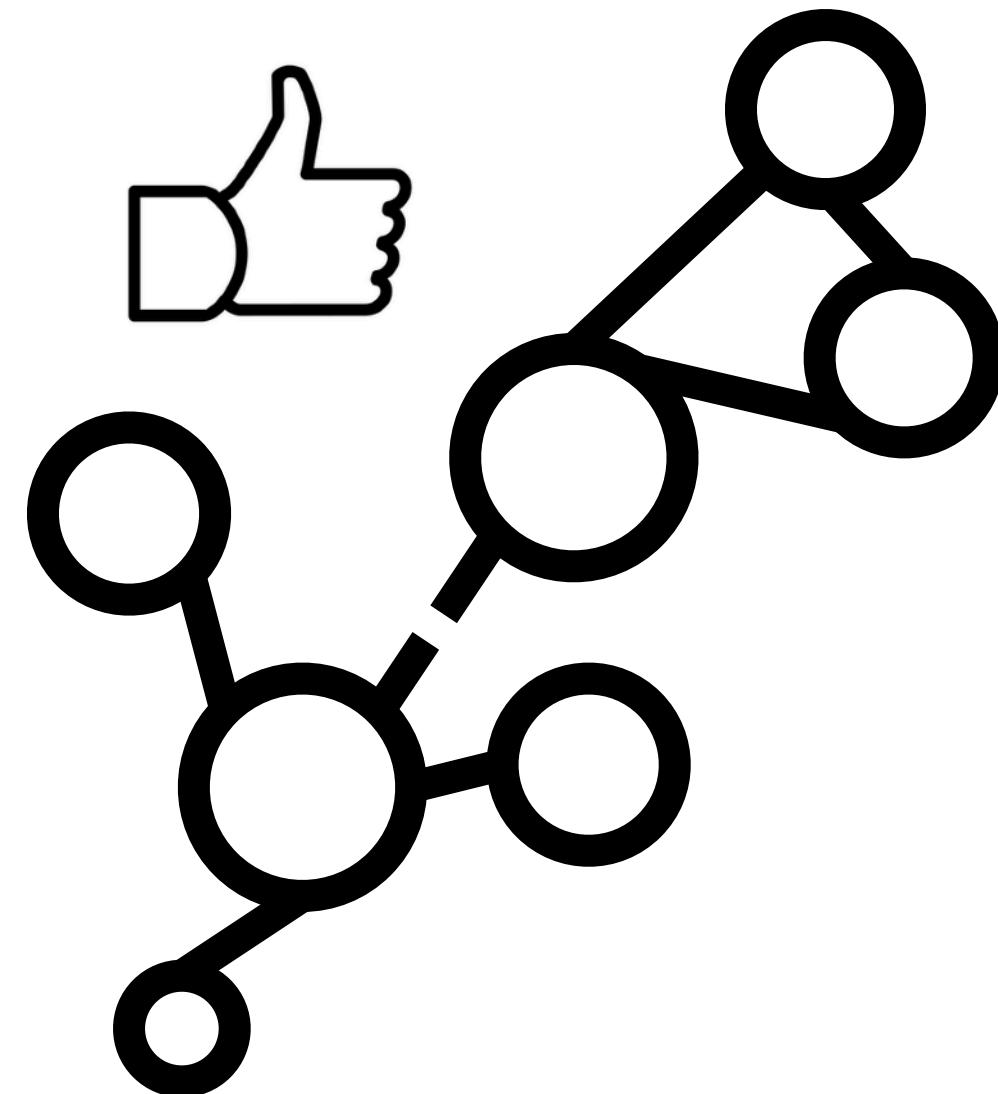


Women in language- (informationally-) diverse teams disengage at lower rates

Survival difference between contributors with high and low language diversity



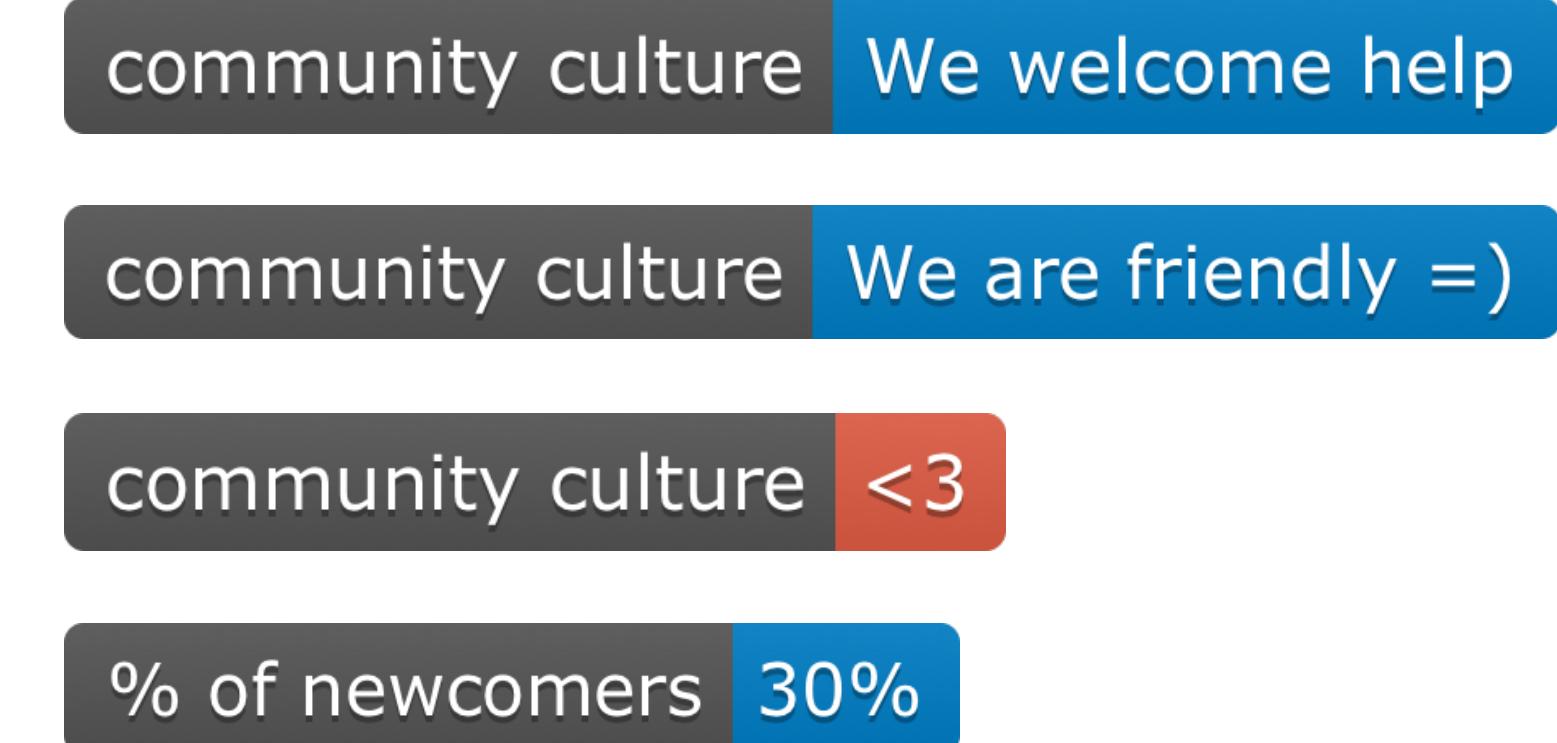
Take away: Invest in building social capital & Foster informationally diverse teams



Recommend projects that can help build social capital



Find relevant mentorship



Signal social capital moderators

Creating sustainable open source communities is hard

Maybe even harder today than ever before
... because of how open source has changed



Today: more problems than solutions

Carnegie Mellon University School of Computer Science STRUDEL

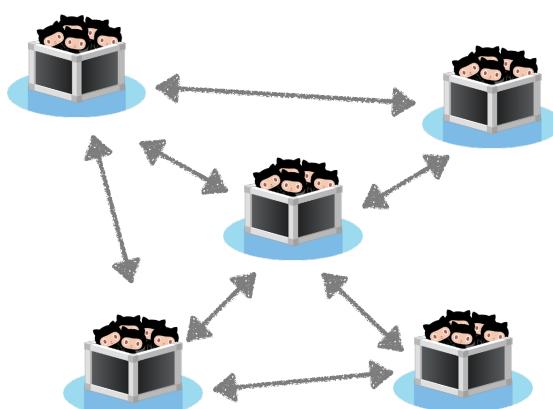
6

Three examples

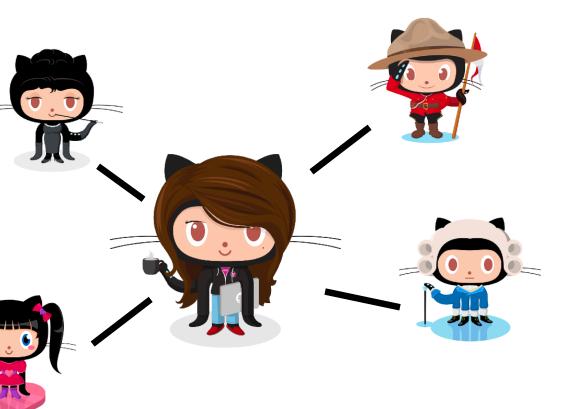
Leveraging transparency

[codementor](https://codementor.io/) GET HELP ▶ npm v1.0.0 Donate
build passing code style standard tips \$3.64/week
ember observer 8 / 10 dependencies out of date
bitHound passing license BSD vulnerabilities 0
bitHound 97 Star 4k cdnjs v3.2.1
bower v3.1.4 downloads 654/month codacy A
issue resolution 3 h coverage 53% Follow 350
slack join build passing dependencies insecure
dependencies up to date IRC irc.freenode.net/unshift
coverage 94% gitter join chat code climate 4.0
build passing commitizen friendly version 4.2.1
release v2.1.1 Greenkeeper enabled docs
Patreon slack 6/160 made by Protocol Labs
code style standard tips \$1.45/week Deploy to Heroku
Forks 847 semantic-release PRs welcome

Considering the whole ecosystem



Building social capital

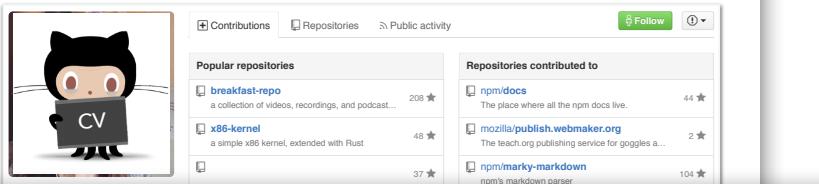


Carnegie Mellon University School of Computer Science STRUDEL

22

Change #3: High level of transparency

- Profile pages for users and projects



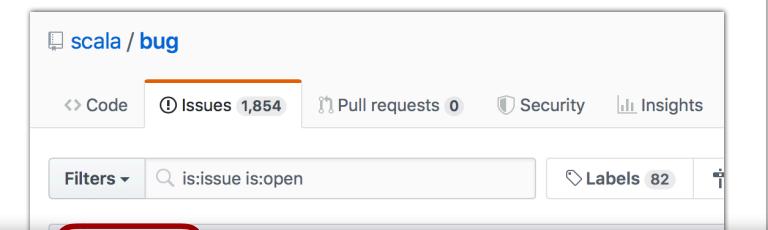
- Rich inference of expertise and interests

- Impacts communication, recruiting and hiring

▷ (Dabbish et al., 2014)
(Marlow et al., 2014)

- Easy to report issues / submit PRs

▷ Growing volume of requests



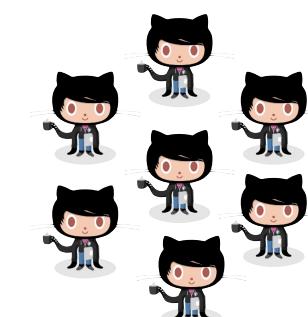
- Social pressure to respond quickly

▷ Otherwise, off-line (Steinmacher et al., 2014)

- Entitlement, unreciprocated

▷ "I have been waiting for 'progress' even though I have done nothing."
▷ "Thank you for your support!"

- Gender representation reality



- Expectation



Change #7: High level of demands & stress

- Easy to report issues / submit PRs

▷ Growing volume of requests

- Social pressure to respond quickly

▷ Otherwise, off-line (Steinmacher et al., 2014)

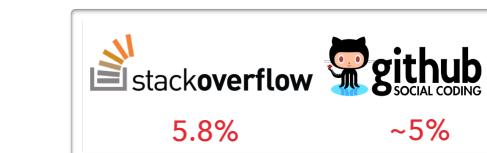
- Entitlement, unreciprocated

▷ "I have been waiting for 'progress' even though I have done nothing."
▷ "Thank you for your support!"

- Gender representation reality

Change #8: Low demographic diversity

- Expectation



"More about the contributions to the code than the 'characteristics' of the person"

"Any demographic identity is irrelevant"

"Code sees no color or gender"

- FLOSS 2013: A survey dataset about free software contributors:
challenges for curating, sharing, and combining G Robles, L Arjona-Reina, B Vasilescu, A Serebrenik, JM Gonzalez-Barahona, MSR 2014
- Stack Overflow 2015 Developer Survey (26,086 people from 157 countries)
Alyssa Frazee: <http://datasafari.com/paper-and-pain-in-code.html>
- Inside Microsoft (2015) <https://go.microsoft.com/fwlink/?linkid=845335>
- Exploring the data on gender and GitHub repo ownership
Alyssa Frazee: <http://datasafari.com/paper-and-pain-in-code.html>
- Stack Overflow 2015 Developer Survey (26,086 people from 157 countries)
<http://stackoverflow.com/research/developer-survey-2015/profile-gender>

• Perceptions of Diversity on GitHub: A User Survey. Vasilescu, B., Filkov, V., and Serebrenik, A. CHASE 2015

19

Bogdan Vasilescu

@b_vasilescu

vasilescu@cmu.edu

<http://cmustrudel.github.io>

What are the main
sustainability
challenges to the
open-source projects
you participate in?



Bogdan Vasilescu
@b_vasilescu

vasilescu@cmu.edu
<http://cmustrudel.github.io>