



General Catalyst x Brandeis

A Glance at Trending Technologies

Xin Yao, Xi Qian, Giang Nguyen

Introduction

- Three team members worked on different but connected parts and mined data as much as possible.
- The project officially started late January and ended yesterday

Objectives

- Modern web and mobile applications are composed of many underlying technologies and services from a variety of vendors
- The relative popularity of these components is of interest to a variety of audiences, including developers and investors.
- The goal for this project is to research, test and prototype several discovery tools.

Methodology

Technology WhiteList

- Database, Sales automation, Hosting, CMS
- Focus: Angular, Ionic, React, JQuery, Hubspot

Methodology

- Mining data from different sources
- Normalize our findings
- Scoring people and their techs!

Data Sources

1. CommonCrawl
2. GitHub
3. Reddit / HackerNews

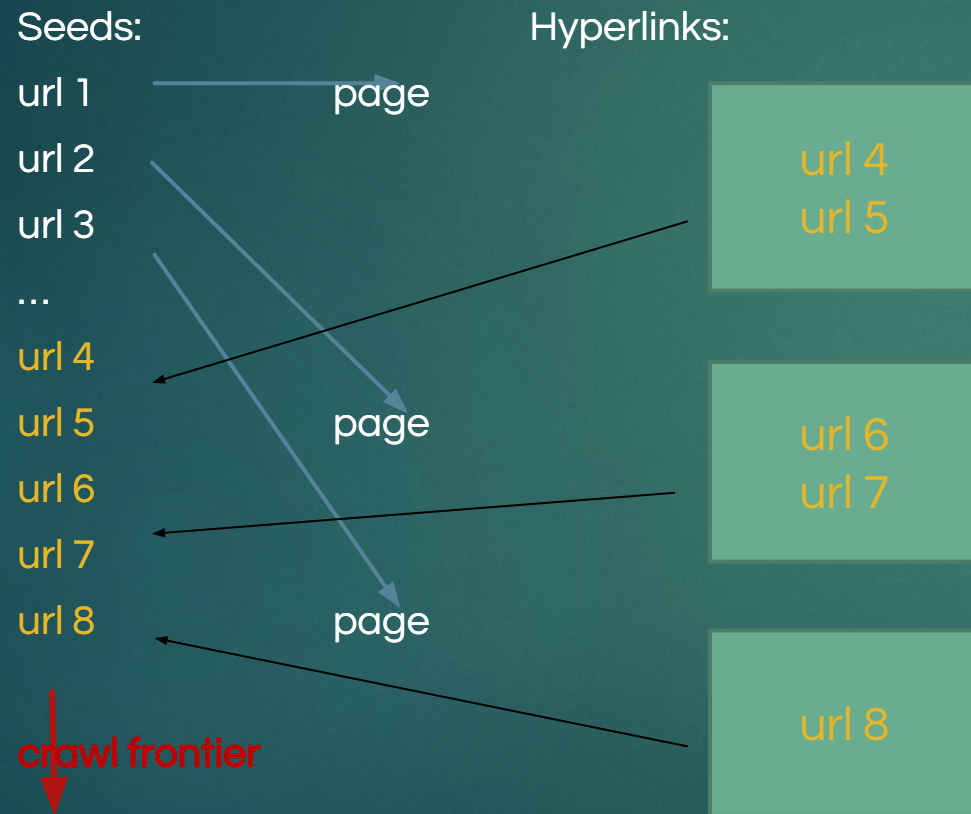
Sources that we either did not successfully utilize or did not dig deep enough:

- LinkedIn
- Facebook / Twitter / Vk / Weibo...
- Proprietary resources (Alexa, builtwith,...)

What is Common Crawl

- It is a nonprofit organization that crawls the web and freely provides its archives and datasets to the public.
- The Web Data Commons project extracts HTML page data from several billion web pages.
- Common crawl's data is stored on Amazon's S3 service, and developed by people from google group. (<http://commoncrawl.org/about/team/>)
- They use web crawler which starts with a list of URLs to visit, and visits these URLs, identifies all the hyperlinks in the page and adds them to the list of URLs to visit.

Web Crawler



How big is Common Crawl?

- The data before November 2013 is Arc file format.
- The data for each month is huge, like really huge, hundreds-of-TBs-per-month huge
- There are total 18-month datasets for both 2014 and 2015.

Crawl Date	Availability date	Size in TB	Billion of pages
November 2015	December 2015	151	1.82
September 2015	November 2015	106	1.32
July 2015	August 2015	145	1.81
June 2015	July 2015	131	1.67
May 2015	July 2015	159	2.05
April 2015	May 2015	168	2.11
March 2015	May 2015	124	1.64
February 2015	March 2015	145	1.9
January 2015	March 2015	139	1.82
December 2014	January 2015	160	2.08
November 2014	December 2014	135	1.95
October 2014	November 2014	254	3.7
September 2014	November 2014	220	2.8
August 2014	September 2014	200	2.8
July 2014	August 2014	266	3.6
April 2014	July 2014	183	2.6
March 2014	March 2014	223	2.8
January 2014	January 2014	148	2.3

Data format

1. Common crawl currently stores the crawl data using the Web Archive (WARC) format. Before that point, the crawl was stored in the ARC file format.
2. WAT file format stores computed metadata for the data stored in WARC files.
3. WET file format stores only extracted plaintext from the data stored in WARC files.

WARC file format

- Not only does the format store the HTTP response from the websites it contacts, it also stores information about how that information was requested and metadata of the process..
- In the example below, we can see the crawler contacted <http://news.bbc.co.uk/2/hi/africa/> and received a HTML page in response. We can also see the page was served from the Apache web server, sets caching details, and attempts to set a cookie

```
HTTP/1.1 200 OK
Server: Apache
Vary: X-CDN
Cache-Control: max-age=0
Content-Type: text/html
Date: Sat, 02 Aug 2014 09:52:13 GMT
Expires: Sat, 02 Aug 2014 09:52:13 GMT
Connection: close
Set-Cookie: BBC-UID=...; expires=Sun, 02-Aug-15 09:52:13 GMT; path=/; domain=bbc.co.uk;

<!doctype html public "-//W3C//DTD HTML 4.0 Transitional//EN" "http://www.w3.org/TR/REC-html40/loose.dtd">
<html>
<head>
<title>
    BBC NEWS | Africa | Namibia braces for Nujoma exit
</title>
...
```


WAT file format

- WAT files contains important metadata stored in the WARC file format above. This metadata is computed for each of the three types of records. If the information crawled is HTML, the computed metadata includes the HTTP headers returned and links listed on the page.
- WAT files are stored as JSON.

```
Envelope
  WARC-Header-Metadata
    WARC-Target-URI [string]
    WARC-Type [string]
    WARC-Date [datetime string]
    ...
  Payload-Metadata
    HTTP-Response-Metadata
      Headers
        Content-Language
        Content-Encoding
        ...
      HTML-Metadata
        Head
          Title [string]
          Link [list]
          Metas [list]
          Links [list]
        Headers-Length [int]
        Entity-Length [int]
        ...
    ...
  ...
```

WET file format

- WET files only contains extracted plaintext.

```
BBC NEWS | Africa | Namibia braces for Nujoma exit
```

```
...
```

```
President Sam Nujoma works in very pleasant surroundings in the small but beautiful old State House...
```


Data mining

1. Web URLs and web IP addresses from WAT files.
2. Technical related information in HTML tags from WARC files.
3. “Snapshots” of websites

Web information

- The data accessed is only 0.1% of the whole common crawl datasets.
- This example contains my extracted data from January 2015. There are totally 17132 web URLs with the corresponding number of visited times by crawler for that month.

web address	number
http://m.mlb.com	41
http://www.atgstores.com	31
http://www.cnet.com	25
http://www.popsugar.com	25
http://www.law.cornell.edu	24
http://www.urbandictionary.com	24
http://www.bloomberg.com	23
http://en.wikipedia.org	23
http://www.agoda.com	23
http://www.summitpost.org	19
http://www.ccc.edu	19
http://www.twopeasinabucket.com	18
http://sourceforge.net	18
http://www.utsandiego.com	18
http://www.timeanddate.com	18
http://www.dailytech.com	18
http://www.dollartree.com	18
http://www.oyster.com	18
http://www.appszoom.com	18
http://www.crunchyroll.com	18
http://www.orientaltrading.com	18
http://www.beau-coup.com	17
http://www.tripadvisor.com	17
http://www.worldcat.org	17
http://www.indeed.com	17
http://www.bccradsports.com	17

Technologies

- This csv file contains all the technologies from web pages that were crawled by crawler in Jan 2015.
- Each corresponding number of technology is based on the number of that web page has been crawled.
- Merged all technologies into 5 categories.

technologies	number
jquery.doubletap.js	14
jquery.fancybox-1.3.1.css	41
angularmods\modules\header_savedlocations\hea	5
jquery/dimensions.js	1
jquery.dateSelector.js	1
jquery/1.3.1/jquery.min.js	23
jquery/js/jquery/cloud-zoom.1.0.2.min.js	3
jquery/js/jquery-1.8.2.js	1
angular_quiz\js\jquery.effects.core.js	1
jquery/jquery.lightbox-0.5.min.js	1
jquery/authforms.js	2
jquery.onImagesLoad.min.js	2
jquery/plugins/xdr.js	6
angularmods\modules\breaking_now\breaking_now	5
jquery.tools.1.2.5.min.js	2
jquery.ddslick.min.js	6
jquery.alignToSpacer.js	2
jquery.form/3.24/jquery.form.min.js	1
jqueryui/1.10.0/themes/base/jquery-ui.css	2
hubspot.net/hub/-1/hub_generated/style_manager/1	2
jquery.nivo.slider.js	34
jquery.infobulle.js	1
jquery.mobile-1.0a4.1/jquery.mobile-1.0a4.1.min.	6
jquery-innerfade-min.js	1
jquery.cycletwo.js	1
jquery.tools.scrollable.css	1

technologies	number
angular	1536
ionic	4
react	86
hubspot	107
jquery	97865

Run our script!



In the future

- Use common crawl **index** and query **api**. It started from late 2015.
- Use Python instead of Java, easier for big data processing syntax wise and availability of packages
- Try other data corpus (Internet Archive, StackOverflow,...)

GitHub

- GitHub: A web-based Git repository hosting service
- Founded and launched in 2008
- Largest host of source code in the world: As of April 2016, GitHub has more than 14 million users and more than 35 million repositories.

Trending on GitHub

Trending in open source

See what the GitHub community is most excited about this week.

Repositories

Developers

Trending: **this week** ▾

All languages

Unknown languages

CSS

HTML

Java

JavaScript

PHP


Python

Ruby

⋮ Other: Languages ▾

ProTip! Looking for most forked repositories? [Try this search](#)


open-power-workgroup/Hospital

Python • 4,020 stars this week • Built by 

★ Star

FreeCodeCamp/FreeCodeCamp


The <https://FreeCodeCamp.com> open source codebase and curriculum. Learn to code and help nonprofits.

JavaScript • 3,947 stars this week • Built by 

★ Star

google/flexbox-layout


Flexbox for Android

Java • 1,932 stars this week • Built by 

★ Star

typicode/json-server

Get a full fake REST API with zero coding in less than 30 seconds (seriously)

JavaScript • 1,826 stars this week • Built by 

★ Star

<https://github.com/trending?since=weekly>

Attempts to explore GitHub data

Score by programming language?

- 35 million repos in less than 100 languages
- Feels like judging people by their zodiac sign
- Try: `language_stats.py`

A glance at the language statistics: using a sample of 1000 repos

Language	Lines	Repos
Ruby	82,219,750	659
JavaScript	30,280,822	281
Shell	2,197,140	101
Python	7,656,529	87
C	49,336,329	85
...
Visual Basic	1724	1
Racket	1172	11
PowerShell	212	1

Attempts to explore Github data

Analyzing metadata of repos?

- 35 million repos, more than 130GB even if we only needed the metadata

Use GitHub datadump?

- Available at <https://www.githubarchive.org/>
- Outdated
- May not contain fields we needed

Attempts to explore Github data

Use GitHub API to build our own database?

- Raw data is not necessarily relational
- Size is still a problem
- API rate limiting

Scoring a Repo

- Preliminary score: Stars + Forks
- Final score: Karma Score * Activeness Score
 - Karma Score: $\log_{10}(\text{Reddit Score}) + \log_{10}(\text{HackerNews Score})$
 - Activeness Score: Commits + Forks + Issue Events with geometric weights in time
 - Commits: Author's activeness
 - Forks: Community's interest
 - Issue Events: Attention and interactions from users

Demo

- Search for repos with live query to GitHub Search API
- Get history data of the repo you are interested in
- The script generates graphs for you


Future usages

- Incorporate more information in a repo's commit / fork / issue event history to provide more accurate scoring
- Explore the network of GitHub users
- Live tracking of all concerned repos, if we had enough computing power

Reddit

MY SUBREDDITS

FRONT - ALL - RANDOM | ASKREDDIT - FUNNY - TODAYILEARNED - PICS - GIFS - GAMING - VIDEOS - WORLDNEWS - NEWS - MOVIES - AWW - SHOWERTHOUGHTS - MILDLYINTERESTING - JOKES - LIFEPROTIP! MORE »

 **reddit**

hot new rising controversial top gilded wiki promoted

Want to join? Log in or sign up in seconds. | English

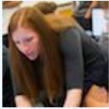
search

username

password

☐ remember me reset password login

↑ 23 ↓



Teach high school CS and keep your day job (c.tealsk12.org)
promoted by TEALS_VOLUNTEERS
49 comments share


sponsored link what's this?

✓ trending subreddits

/r/shockwaveporn /r/unlikelyfriends /r/TheOnion /r/holdmyjuicebox /r/fuckolly 52 comments

↑ 1 ↓


6261



Men who can tell a good story are seen as more attractive and higher status Psychology
(digest.bps.org.uk)
submitted 4 hours ago by Lightfiend to /r/science
617 comments share

↑ 2 ↓


5337



The best Ask Amy response I've ever read (i.imgur.com)
submitted 5 hours ago by drdoom to /r/pics
573 comments share

↑ 3 ↓


5625



President Obama signs bill declaring the bison the national mammal (abc10.com)
submitted 6 hours ago by Another-Chance to /r/news
932 comments share

↑ 4 ↓


5245



Dating website to match Canadians with Americans escaping Trump presidency (cp24.com)
submitted 6 hours ago by TomSawyer_ to /r/nottheonion
1854 comments share

↑ 5 ↓


6142



For everyone who wanted to see the actual demolition (i.imgur.com)
submitted 7 hours ago by Thatdude283 to /r/gifs
879 comments share

↑ 6 ↓


6453



A bobcat and its bobkitten. (i.imgur.com)
submitted 8 hours ago by MaryBoyd to /r/aww
469 comments share

↑ 7 ↓

6334



Nothing to see here (i.imgur.com)



discuss this ad on reddit

Submit a new link

Submit a new text post

Reddit

But there's more to it..


5109




Transport

Uber and Lyft pull out of Austin after locals vote against self-regulation | Technology

(theguardian.com)

submitted 17 hours ago by Christianpaul

3213 comments share


487


Business

Jury is picked for \$9 billion Oracle v. Google showdown. Only one juror worked with computers, and he was Oracle's first strike. (arstechnica.com)

submitted 4 hours ago by redditor_1234

66 comments share

Technology

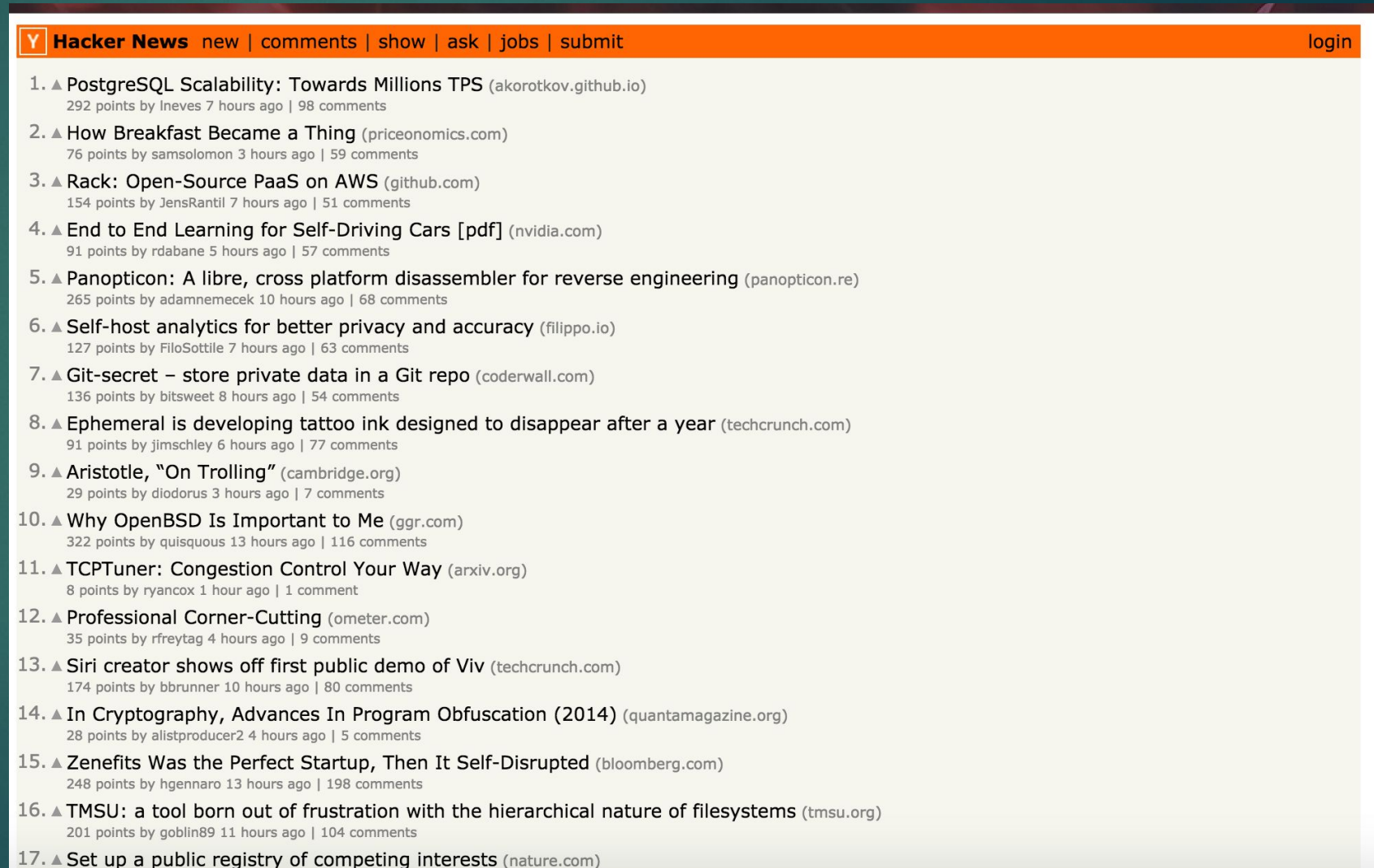
subscribe

5,025,003 readers
● 3,520 users here now

Join the live chat on IRC

HackerNews

- Also known as Ycombinator



The screenshot shows the Hacker News website interface. At the top, there is an orange navigation bar with the 'Y' logo, the text 'Hacker News', and links for 'new', 'comments', 'show', 'ask', 'jobs', and 'submit'. A 'login' link is located on the far right. Below the navigation bar, a list of 17 items is displayed, each starting with a number and an upward arrow. Each item includes a title, a source in parentheses, and a line of meta-information showing the number of points, the user's name, the time since posting, and the number of comments.

Hacker News new | comments | show | ask | jobs | submit login

1. ▲ PostgreSQL Scalability: Towards Millions TPS (akorotkov.github.io)
292 points by Ineves 7 hours ago | 98 comments
2. ▲ How Breakfast Became a Thing (priceonomics.com)
76 points by samsolomon 3 hours ago | 59 comments
3. ▲ Rack: Open-Source PaaS on AWS (github.com)
154 points by JensRantil 7 hours ago | 51 comments
4. ▲ End to End Learning for Self-Driving Cars [pdf] (nvidia.com)
91 points by rdabane 5 hours ago | 57 comments
5. ▲ Panopticon: A libre, cross platform disassembler for reverse engineering (panopticon.re)
265 points by adamnemecek 10 hours ago | 68 comments
6. ▲ Self-host analytics for better privacy and accuracy (filippo.io)
127 points by FiloSottile 7 hours ago | 63 comments
7. ▲ Git-secret – store private data in a Git repo (coderwall.com)
136 points by bitsweet 8 hours ago | 54 comments
8. ▲ Ephemeral is developing tattoo ink designed to disappear after a year (techcrunch.com)
91 points by jimschley 6 hours ago | 77 comments
9. ▲ Aristotle, "On Trolling" (cambridge.org)
29 points by diodorus 3 hours ago | 7 comments
10. ▲ Why OpenBSD Is Important to Me (ggr.com)
322 points by quisquous 13 hours ago | 116 comments
11. ▲ TCPTuner: Congestion Control Your Way (arxiv.org)
8 points by ryancox 1 hour ago | 1 comment
12. ▲ Professional Corner-Cutting (ometer.com)
35 points by rfreytag 4 hours ago | 9 comments
13. ▲ Siri creator shows off first public demo of Viv (techcrunch.com)
174 points by bbrunner 10 hours ago | 80 comments
14. ▲ In Cryptography, Advances In Program Obfuscation (2014) (quantamagazine.org)
28 points by alistproducer2 4 hours ago | 5 comments
15. ▲ Zenefits Was the Perfect Startup, Then It Self-Disrupted (bloomberg.com)
248 points by hgennaro 13 hours ago | 198 comments
16. ▲ TMSU: a tool born out of frustration with the hierarchical nature of filesystems (tmsu.org)
201 points by goblin89 11 hours ago | 104 comments
17. ▲ Set up a public registry of competing interests (nature.com)

Reddit / HackerNews

How we used them:

- Breakdown users' karma by subreddit - communities they associate
- Acquire the "karma" they have and normalize it (log base 10) so we can use that for the final scoring
- More to explore

Future usages

- Pinpoint users that we are interested in and keep tracking
- Sentimental analysis of technologies mention
- Draw correlations, heatmaps of communities of users

Advantages & Disadvantages

- Common Crawl
- GitHub
- Reddit/ HackerNews

CommonCrawl

Advantages:

1. **Giant datasets contain enough web information for data mining.**
2. **Different file formats can be used for different purposes.**
3. **It is FREE & PUBLICLY AVAILABLE data.**

Disadvantages:

1. Too big to process.
2. It is hard to figure out how they crawl the data from websites.
3. Can be manipulated by SEO.

GitHub

Advantages:

1. **Open-source projects galore!**
2. **Activeness of projects can be tracked.**
3. **Also FREE & PUBLICLY AVAILABLE data.**

Disadvantages:

1. API limit.
2. Biased against the community who do not use GitHub

Reddit / HackerNews

Advantages:

1. **Social aspect of projects and their owners**
2. **Can also track the communities tech users associate with**
3. **Did we mention FREE & PUBLICLY AVAILABLE?**

Disadvantages:

1. API limit.
2. People do not use same handles everywhere -> hard to track
3. Ethical issues?