《开源软件与社区治理》

第十二讲 开源社区数据分析

主讲人：赵生宇

X-lab
开放实验室

内容

1 / **为什么**
Why

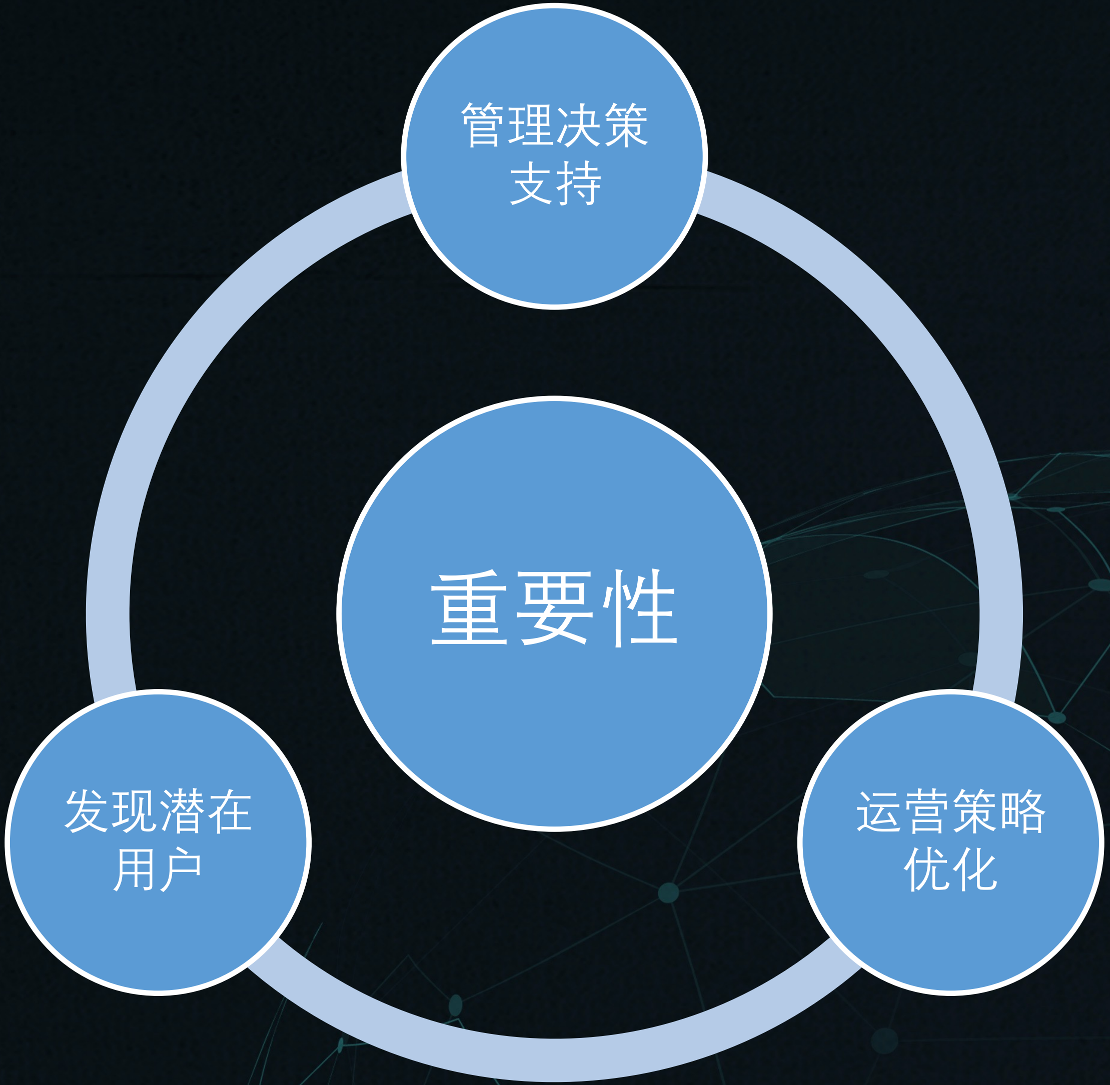
2 / **数据&可视化**
Data&Visualization

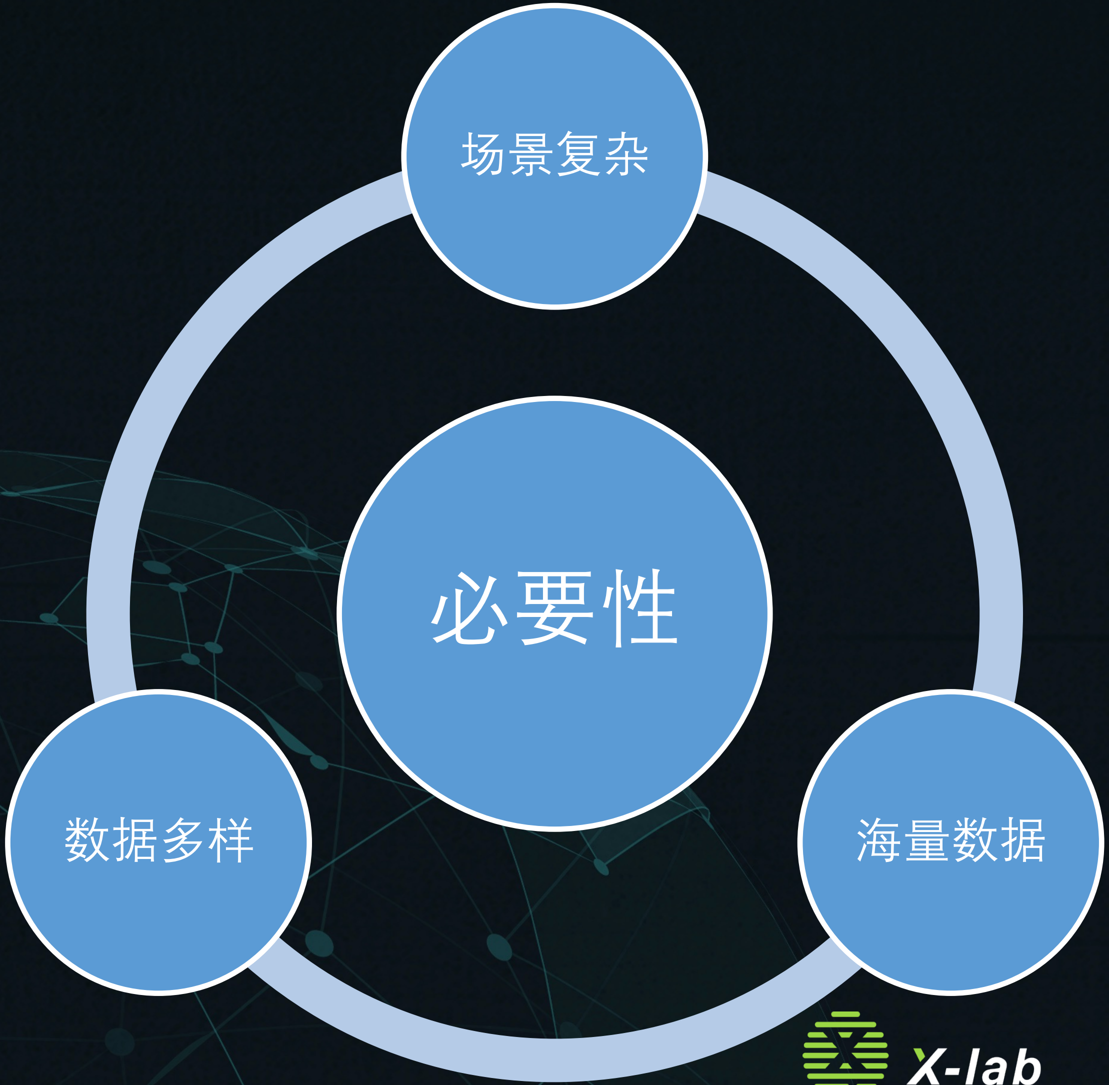
3 / **问题**
Problems

4 / **举个例子**
Example

01 / 为什么

Why

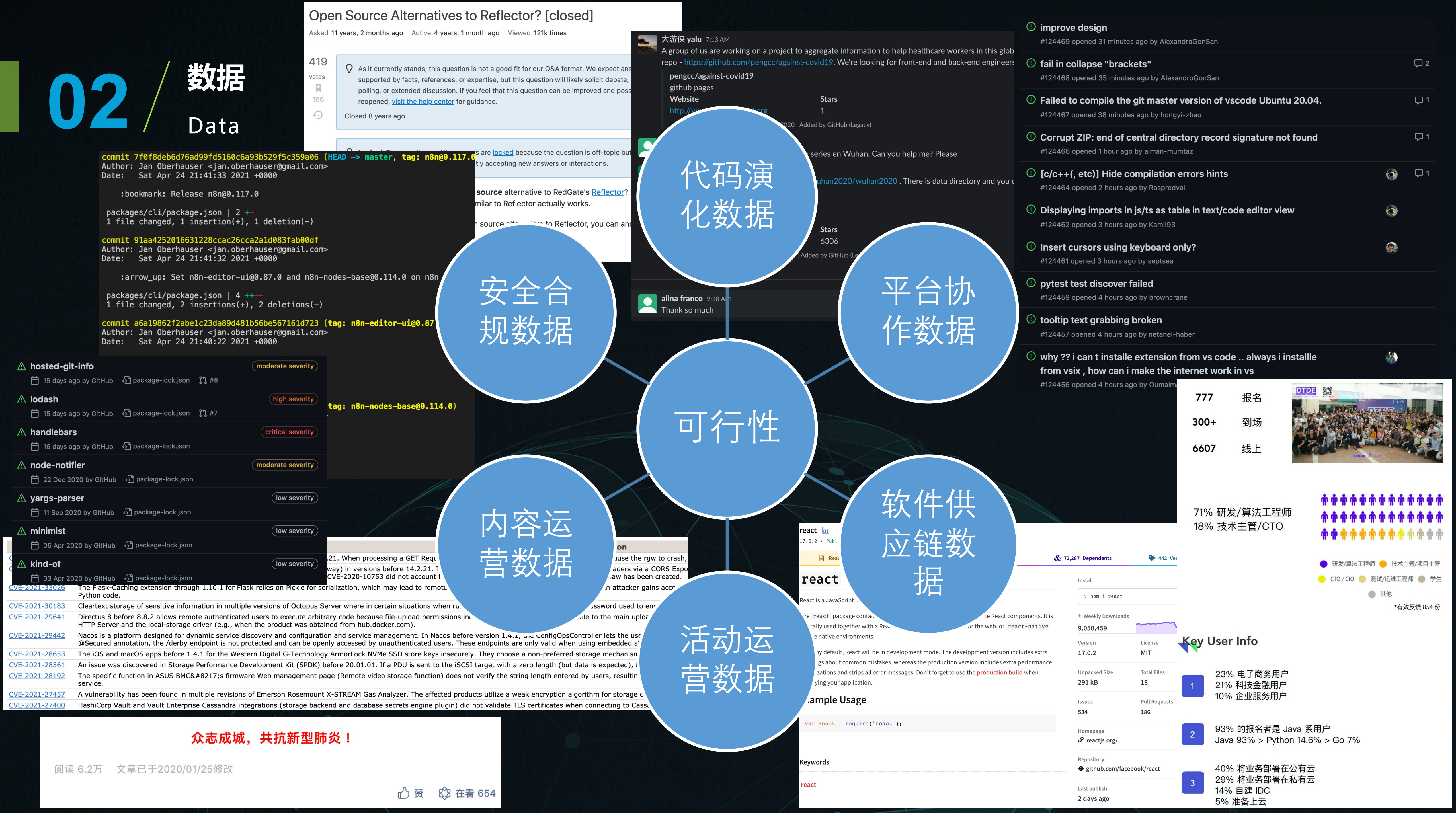




X-lab 开放实验室

02 / 数据

Data



Open Source Alternatives to Reflector? [closed]

Asked 11 years, 2 months ago Active 4 years, 1 month ago Viewed 121k times

419 votes 156

As it currently stands, this question is not a good fit for our Q&A format. We expect ans supported by facts, references, or expertise, but this question will likely solicit debate, polling, or extended discussion. If you feel that this question can be improved and poss reopened, visit the help center for guidance.

Closed 8 years ago.

s are locked because the question is off-topic but tly accepting new answers or interactions.

source alternative to RedGate's Reflector? milar to Reflector actually works.

source alt to Reflector, you can ans

大游侠 yalu 7:13 AM

A group of us are working on a project to aggregate information to help healthcare workers in this glob repo - https://github.com/pengcc/against-covid19. We're looking for front-end and back-end engineers

pengcc/against-covid19

github pages

Website	Stars
http://w org	1

2020 | Added by GitHub (Legacy)

series en Wuhan. Can you help me? Please

uhan2020/wuhan2020 . There is data directory and you c

Stars
6306

Added by GitHub (Le

alina franco 9:18 AM

Thank so much

```
commit 7f0f8deb6d76ad99fd5160c6a93b529f5c359a06 (HEAD -> master, tag: n8n@0.117.0
Author: Jan Oberhauser <jan.oberhauser@gmail.com>
Date:   Sat Apr 24 21:41:33 2021 +0000

    :bookmark: Release n8n@0.117.0

 packages/cli/package.json | 2 +-
 1 file changed, 1 insertion(+), 1 deletion(-)

commit 91aa4252016631228ccac26cca2a1d083fab00df
Author: Jan Oberhauser <jan.oberhauser@gmail.com>
Date:   Sat Apr 24 21:41:32 2021 +0000

    :arrow_up: Set n8n-editor-ui@0.87.0 and n8n-nodes-base@0.114.0 on n8n

 packages/cli/package.json | 4 ++--
 1 file changed, 2 insertions(+), 2 deletions(-)

commit a6a19862f2abe1c23da89d481b56be567161d723 (tag: n8n-editor-ui@0.87
Author: Jan Oberhauser <jan.oberhauser@gmail.com>
Date:   Sat Apr 24 21:40:22 2021 +0000

tag: n8n-nodes-base@0.114.0)
```

improve design
#124469 opened 31 minutes ago by AlexandroGonSan

fail in collapse "brackets" 2
#124468 opened 35 minutes ago by AlexandroGonSan

Failed to compile the git master version of vscode Ubuntu 20.04. 1
#124467 opened 38 minutes ago by hongyi-zhao

Corrupt ZIP: end of central directory record signature not found 1
#124466 opened 1 hour ago by aiman-mumtaz

[c/c++(, etc)] Hide compilation errors hints 1
#124464 opened 2 hours ago by Raspredval

Displaying imports in js/ts as table in text/code editor view
#124462 opened 3 hours ago by Kamil93

Insert cursors using keyboard only?
#124461 opened 3 hours ago by septsea

pytest test discover failed
#124459 opened 4 hours ago by browncrane

tooltip text grabbing broken
#124457 opened 4 hours ago by netanel-haber

why ?? i can t installe extension from vs code .. always i installle from vsix , how can i make the internet work in vs
#124456 opened 4 hours ago by Oumaim

hosted-git-info — moderate severity
15 days ago by GitHub package-lock.json #8

lodash — high severity
15 days ago by GitHub package-lock.json #7

handlebars — critical severity
16 days ago by GitHub package-lock.json

node-notifier — moderate severity
22 Dec 2020 by GitHub package-lock.json

yargs-parser — low severity
11 Sep 2020 by GitHub package-lock.json

minimist — low severity
06 Apr 2020 by GitHub package-lock.json

kind-of — low severity
03 Apr 2020 by GitHub package-lock.json

on

.21. When processing a GET Requ use the rgw to crash,

way) in versions before 14.2.21. T aders via a CORS Expo

CVE-2020-10753 did not account f aw has been created.

CVE-2021-33026	The Flask-Caching extension through 1.10.1 for Flask relies on Pickle for serialization, which may lead to remote n attacker gains acce Python code.
CVE-2021-30183	Cleartext storage of sensitive information in multiple versions of Octopus Server where in certain situations when ru ssword used to enc
CVE-2021-29641	Directus 8 before 8.8.2 allows remote authenticated users to execute arbitrary code because file-upload permissions inc ile to the main uploa HTTP Server and the local-storage driver (e.g., when the product was obtained from hub.docker.com).
CVE-2021-29442	Nacos is a platform designed for dynamic service discovery and configuration and service management. In Nacos before version 1.4.1, the ConfigOpsController lets the use @Secured annotation, the /derby endpoint is not protected and can be openly accessed by unauthenticated users. These endpoints are only valid when using embedded s
CVE-2021-28653	The iOS and macOS apps before 1.4.1 for the Western Digital G-Technology ArmorLock NVMe SSD store keys insecurely. They choose a non-preferred storage mechanism
CVE-2021-28361	An issue was discovered in Storage Performance Development Kit (SPDK) before 20.01.01. If a PDU is sent to the iSCSI target with a zero length (but data is expected),
CVE-2021-28192	The specific function in ASUS BMC&#8217;s firmware Web management page (Remote video storage function) does not verify the string length entered by users, resultin service.
CVE-2021-27457	A vulnerability has been found in multiple revisions of Emerson Rosemount X-STREAM Gas Analyzer. The affected products utilize a weak encryption algorithm for storage c
CVE-2021-27400	HashiCorp Vault and Vault Enterprise Cassandra integrations (storage backend and database secrets engine plugin) did not validate TLS certificates when connecting to Cass

众志成城，共抗新型肺炎！

阅读 6.2万 文章已于2020/01/25修改

赞 在看 654

react DT
17.0.2 • Publ

Rea

72,287 Dependents 442 Ver

react

React is a JavaScript l

e react package conta e React components. It is ally used together with a Rea or the web, or react-native e native environments.

y default, React will be in development mode. The development version includes extra gs about common mistakes, whereas the production version includes extra performance zations and strips all error messages. Don't forget to use the production build when ying your application.

ample Usage

```
var React = require('react');
```

Keywords

react

Install

```
> npm i react
```

Weekly Downloads
9,050,459

Version	License
17.0.2	MIT
Unpacked Size	**Total Files**
291 kB	18
Issues	**Pull Requests**
534	186

Homepage
reactjs.org/

Repository
github.com/facebook/react

Last publish
2 days ago

777	报名
300+	到场
6607	线上

71% 研发/算法工程师
18% 技术主管/CTO

● 研发/算法工程师 ● 技术主管/项目主管 ● CTO / CIO ● 测试/运维工程师 ● 学生 ● 其他

*有效反馈 854 份

Key User Info

1 23% 电子商务用户
21% 科技金融用户
10% 企业服务用户

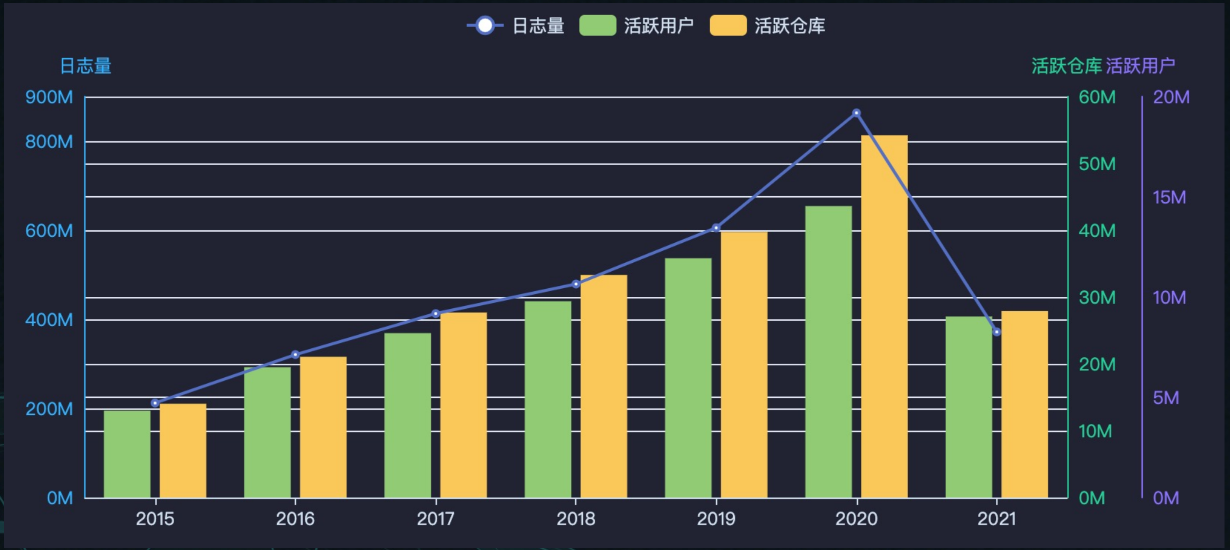
2 93% 的报名者是 Java 系用户
Java 93% > Python 14.6% > Go 7%

3 40% 将业务部署在公有云
29% 将业务部署在私有云
14% 自建 IDC
5% 准备上云

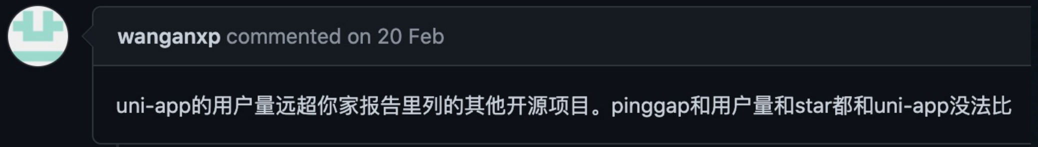
02 数据 Data

GitHub 事件日志

- 2015 至今共计 30 亿+ 条行为日志记录
- 覆盖全域所有项目的所有数据
- 覆盖 star、fork、issue、pull request、release、push、wiki 等主要协作数据

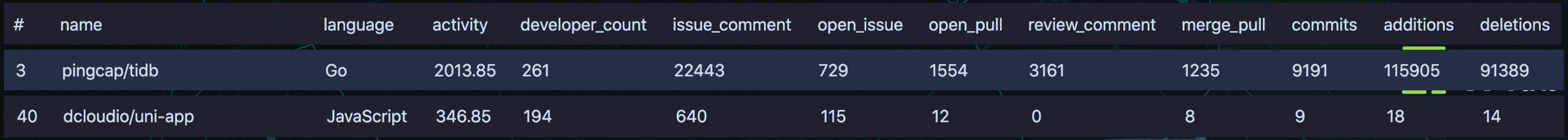


- 用数据说话，才能知道真相
  - GitHub 数据
    - Octoverse 2020：用户总量 5600W，新建仓库 6000W
    - 2020 日志数据：活跃用户 1454W，活跃仓库 5421W
  - TiDB(2W+ star) vs uni-app(3W+ star)



wanganxp commented on 20 Feb

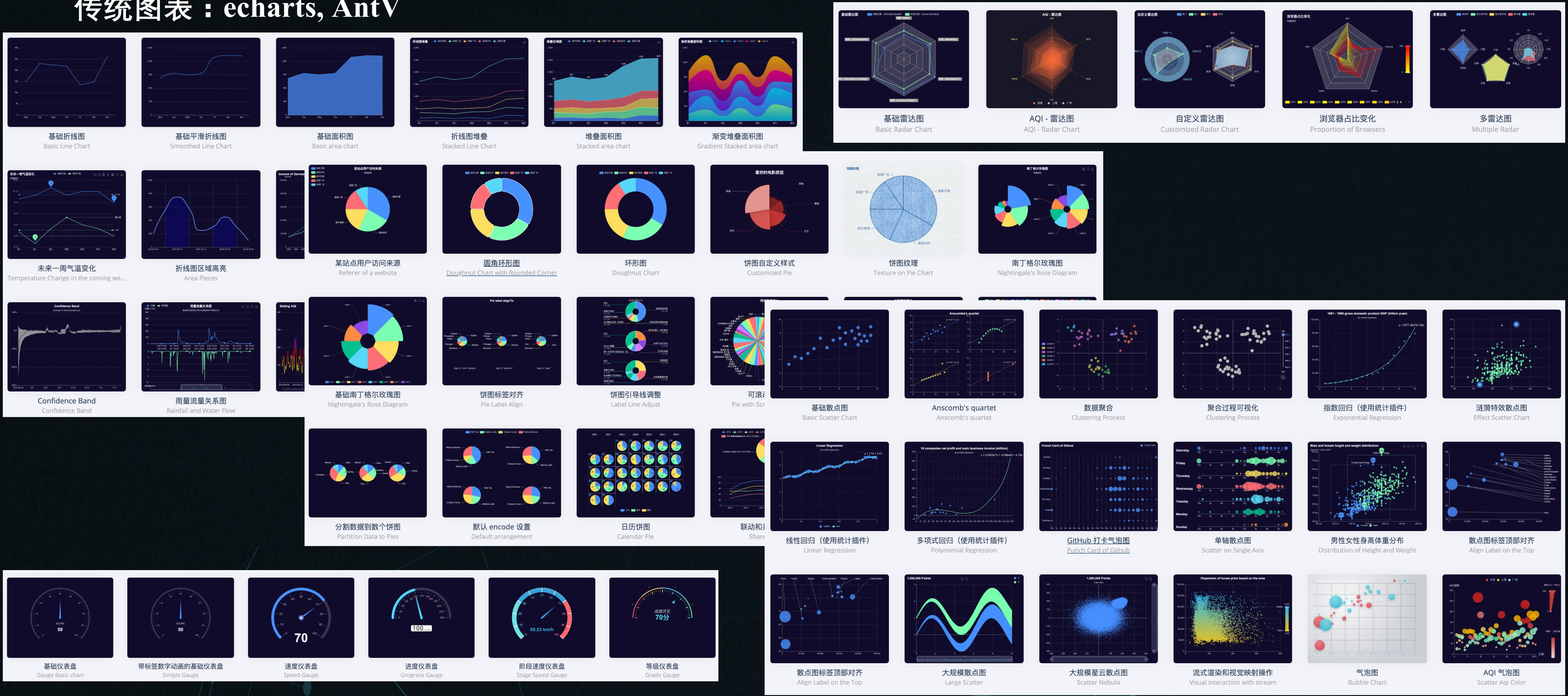
uni-app的用户量远超你家报告里列的其他开源项目。pinggap和用户量和star都和uni-app没法比



#	name	language	activity	developer_count	issue_comment	open_issue	open_pull	review_comment	merge_pull	commits	additions	deletions
3	pingcap/tidb	Go	2013.85	261	22443	729	1554	3161	1235	9191	115905	91389
40	dcloudio/uni-app	JavaScript	346.85	194	640	115	12	0	8	9	18	14

02 可视化 Visualization

传统图表 ：echarts, AntV



基础折线图 Basic Line Chart

基础平滑折线图 Smoothed Line Chart

基础面积图 Basic area chart

折线图堆叠 Stacked Line Chart

堆叠面积图 Stacked area chart

渐变堆叠面积图 Gradient Stacked area chart

未来一周气温变化 Temperature Change in the coming we...

折线图区域高亮 Area Pieces

Confidence Band Confidence Band

雨量流量关系图 Rainfall and Water Flow

基础雷达图 Basic Radar Chart

AQI - 雷达图 AQI - Radar Chart

自定义雷达图 Customized Radar Chart

浏览器占比变化 Proportion of Browsers

多雷达图 Multiple Radar

某站点用户访问来源 Referer of a website

圆角环形图 Doughnut Chart with Rounded Corner

环形图 Doughnut Chart

饼图自定义样式 Customized Pie

饼图纹理 Texture on Pie Chart

南丁格尔玫瑰图 Nightingale's Rose Diagram

基础南丁格尔玫瑰图 Nightingale's Rose Diagram

饼图标签对齐 Pie Label Align

饼图引导线调整 Label Line Adjust

可滚 Pie with Scr

分割数据到数个饼图 Partition Data to Pies

默认 encode 设置 Default arrangement

日历饼图 Calendar Pie

联动和 Share

基础散点图 Basic Scatter Chart

Anscomb's quartet Anscomb's quartet

数据聚合 Clustering Process

聚合过程可视化 Clustering Process

指数回归（使用统计插件） Exponential Regression

涟漪特效散点图 Effect Scatter Chart

线性回归（使用统计插件） Linear Regression

多项式回归（使用统计插件） Polynomial Regression

GitHub 打卡气泡图 Punch Card of Github

单轴散点图 Scatter on Single Axis

男性女性身高体重分布 Distribution of Height and Weight

散点图标签顶部对齐 Align Label on the Top

散点图标签顶部对齐 Align Label on the Top

大规模散点图 Large Scatter

大规模星云散点图 Scatter Nebula

流式渲染和视觉映射操作 Visual interaction with stream

气泡图 Bubble Chart

AQI 气泡图 Scatter Aqi Color

基础仪表盘 Gauge Basic chart

带标签数字动画的基础仪表盘 Simple Gauge

速度仪表盘 Speed Gauge

进度仪表盘 Grogress Gauge

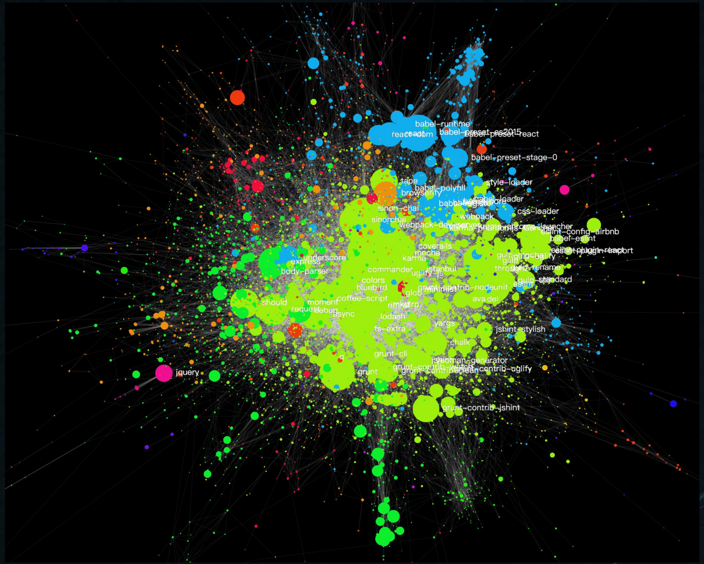
阶段速度仪表盘 Stage Speed Gauge

等级仪表盘 Grade Gauge

02 可视化

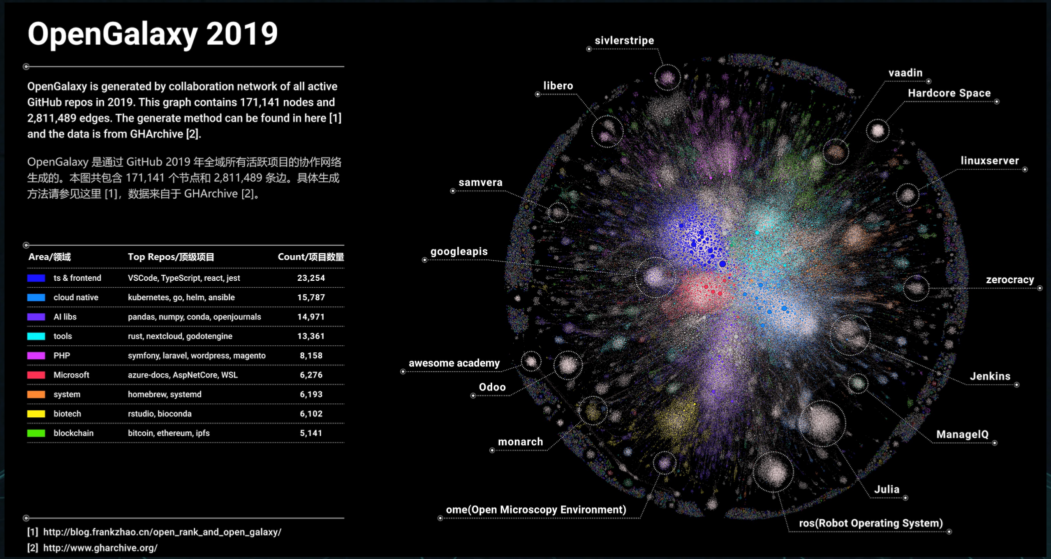
Visualization

图（Graph）类型



npm 包依赖网络

echarts GraphGL
54,000 节点
10,000 边
npm 总量 181W



OpenGalaxy 2019

OpenGalaxy is generated by collaboration network of all active GitHub repos in 2019. This graph contains 171,141 nodes and 2,811,489 edges. The generate method can be found in here [1] and the data is from GHArchive [2].

OpenGalaxy 是通过 GitHub 2019 年全域所有活跃项目的协作网络生成的。本图共包含 171,141 个节点和 2,811,489 条边。具体生成方法请参见这里 [1]，数据来自于 GHArchive [2]。

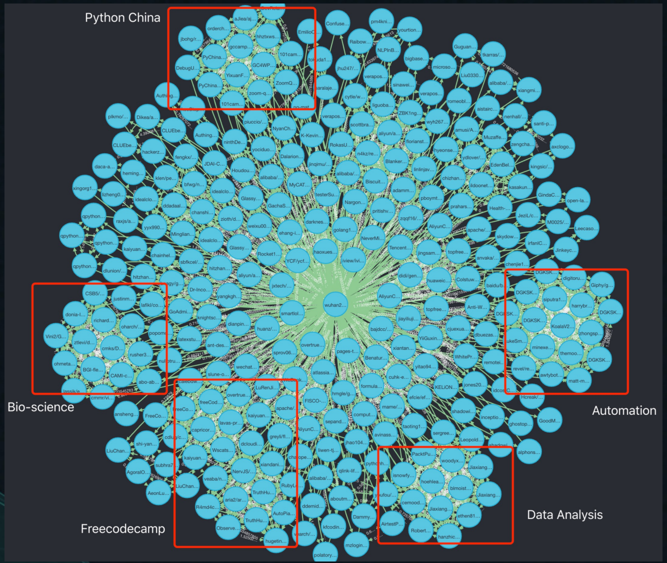
Area/领域	Top Repos/顶级项目	Count/项目数量
🟦 ts & frontend	VSCode, TypeScript, react, jest	23,254
🟦 cloud native	kubernetes, go, helm, ansible	15,787
🟪 AI libs	pandas, numpy, conda, openjournals	14,971
🟦 tools	rust, nextcloud, godotengine	13,361
🟪 PHP	symfony, laravel, wordpress, magento	8,158
🟥 Microsoft	azure-docs, AspNetCore, WSL	6,276
🟧 system	homebrew, systemd	6,193
🟨 biotech	rstudio, bioconda	6,102
🟩 blockchain	bitcoin, ethereum, ipfs	5,141

[1] http://blog.frankzhao.cn/open_rank_and_open_galaxy/

[2] http://www.gharchive.org/

OpenGalaxy 2019

Gephi
171,141 节点
2,811,489 边
GitHub 总量



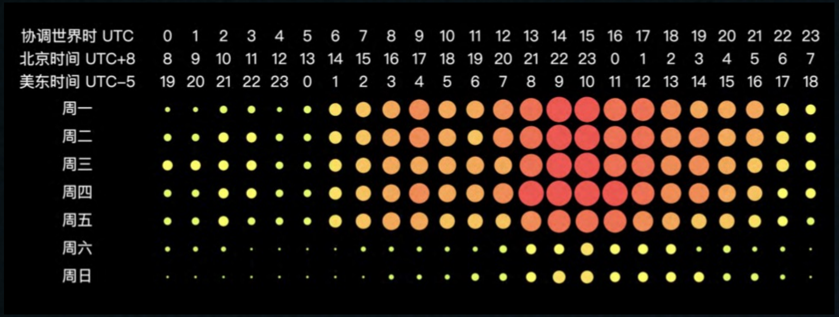
Wuhan2020 协作项目关系图

Neo4j
300 节点
862 边

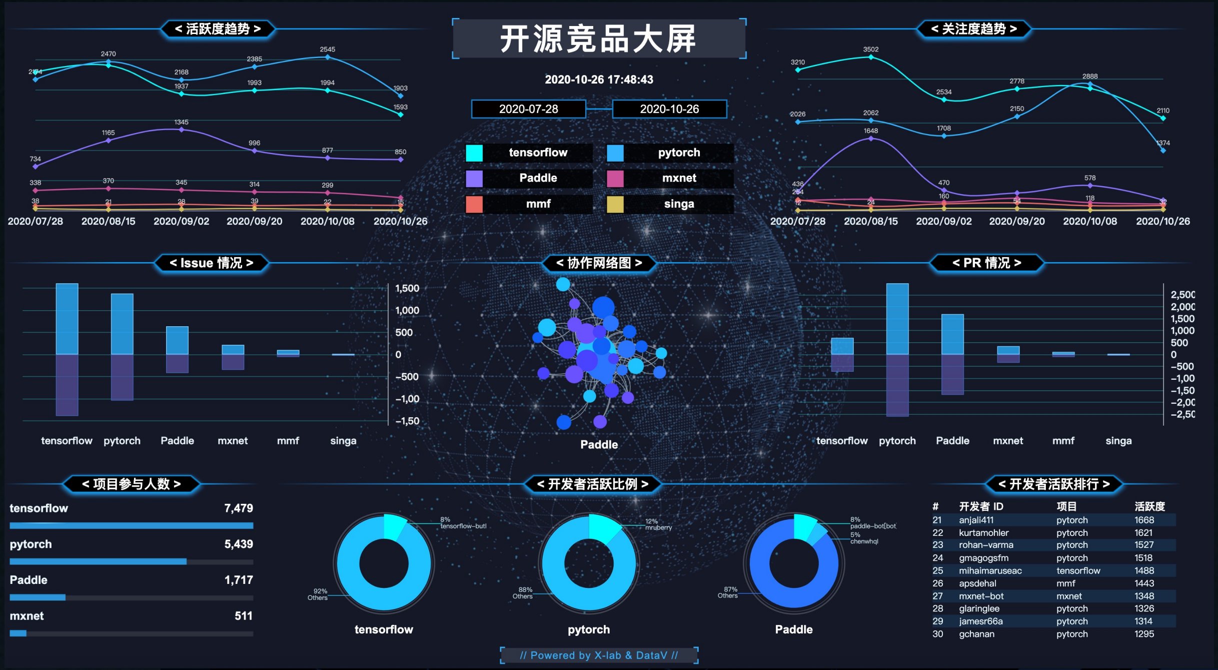
02 / 可视化

Visualization

自定义可视化



日志时间分布打孔图



开源竞品大屏

2020-10-26 17:48:43

2020-07-28 — 2020-10-26

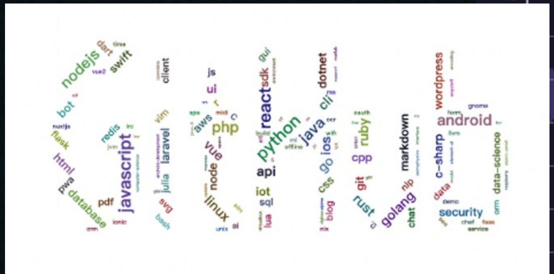
< 项目参与人数 >

tensorflow	7,479
pytorch	5,439
Paddle	1,717
mxnet	511

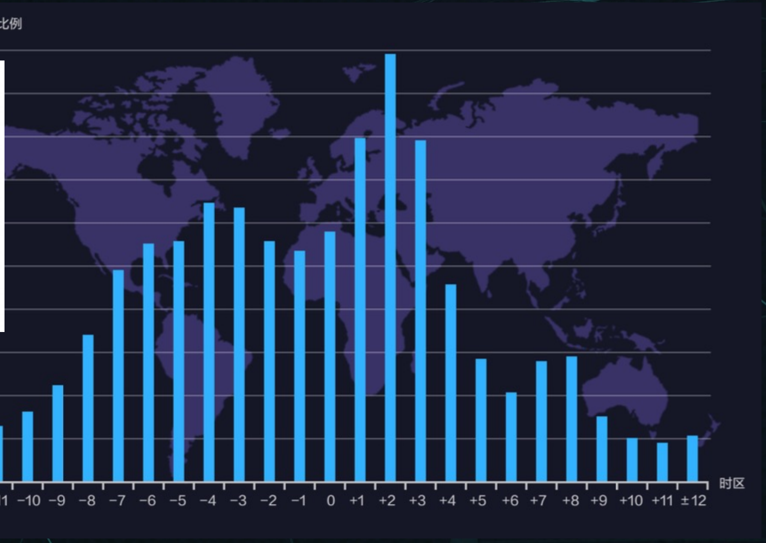
< 开发者活跃排行 >

#	开发者 ID	项目	活跃度
21	anjali411	pytorch	1668
22	kurtamohler	pytorch	1621
23	rohan-varma	pytorch	1527
24	gmagogsfm	pytorch	1518
25	mihaimaruseac	tensorflow	1488
26	apsdehal	mmf	1443
27	mxnet-bot	mxnet	1348
28	glaringlee	pytorch	1326
29	jamesr66a	pytorch	1314
30	gchanan	pytorch	1295

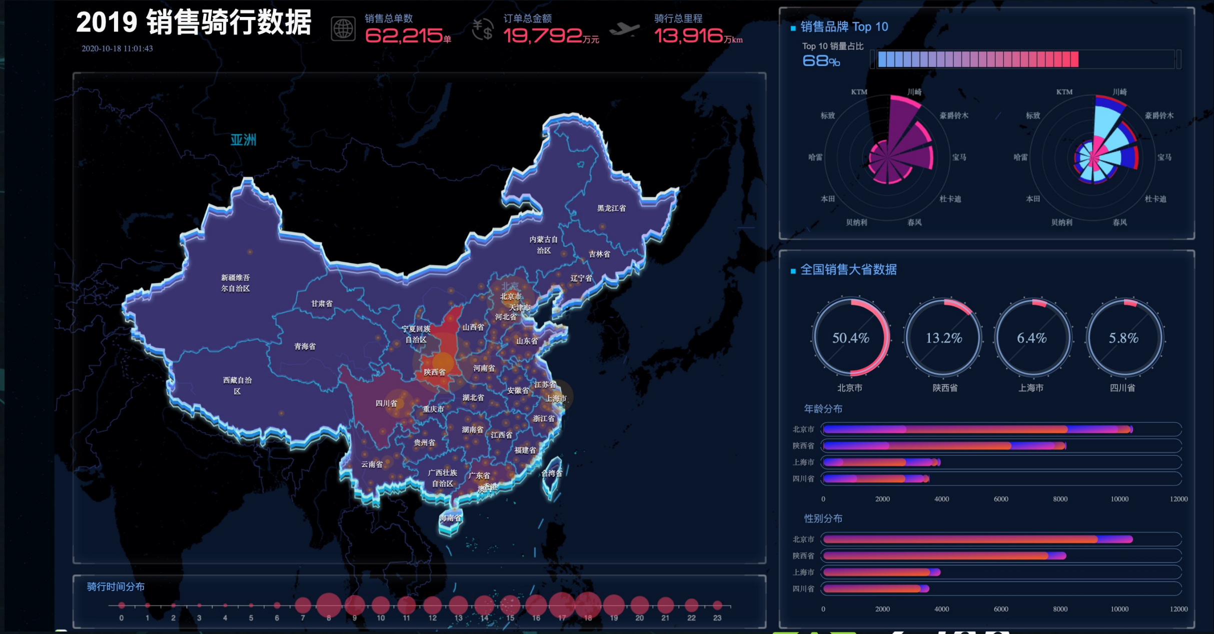
// Powered by X-lab & DataV //



词云



人员分布时区柱状图



2019 销售骑行数据

2020-10-18 11:01:43

销售总单数 62,215单

订单总金额 19,792万元

骑行总里程 13,916万km

聚合数据大屏

http://oss.x-lab.info/github-insight-report-2020.pdf

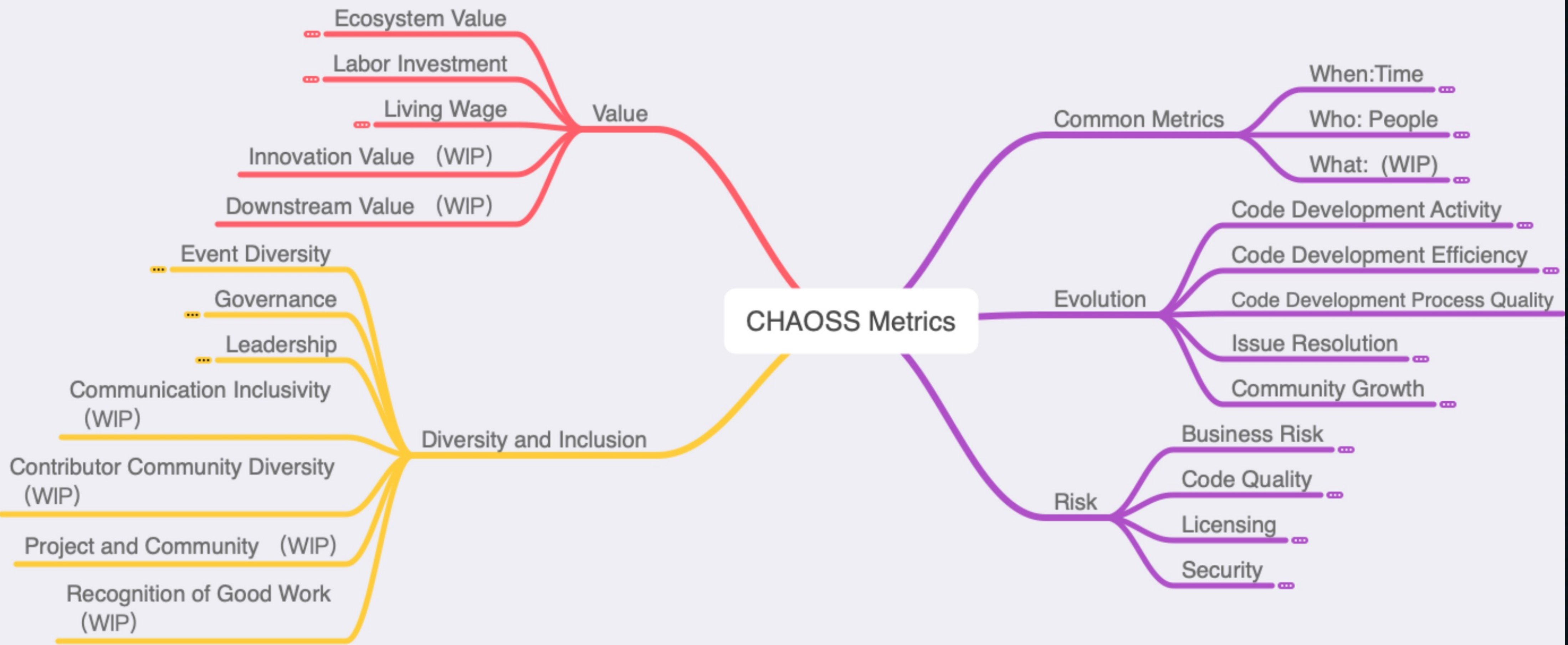
http://oss.x-lab.info/github-insight-report-2020-en.pdf

X-lab 开放实验室

03 问题

Problems

指标体系——项目健康度



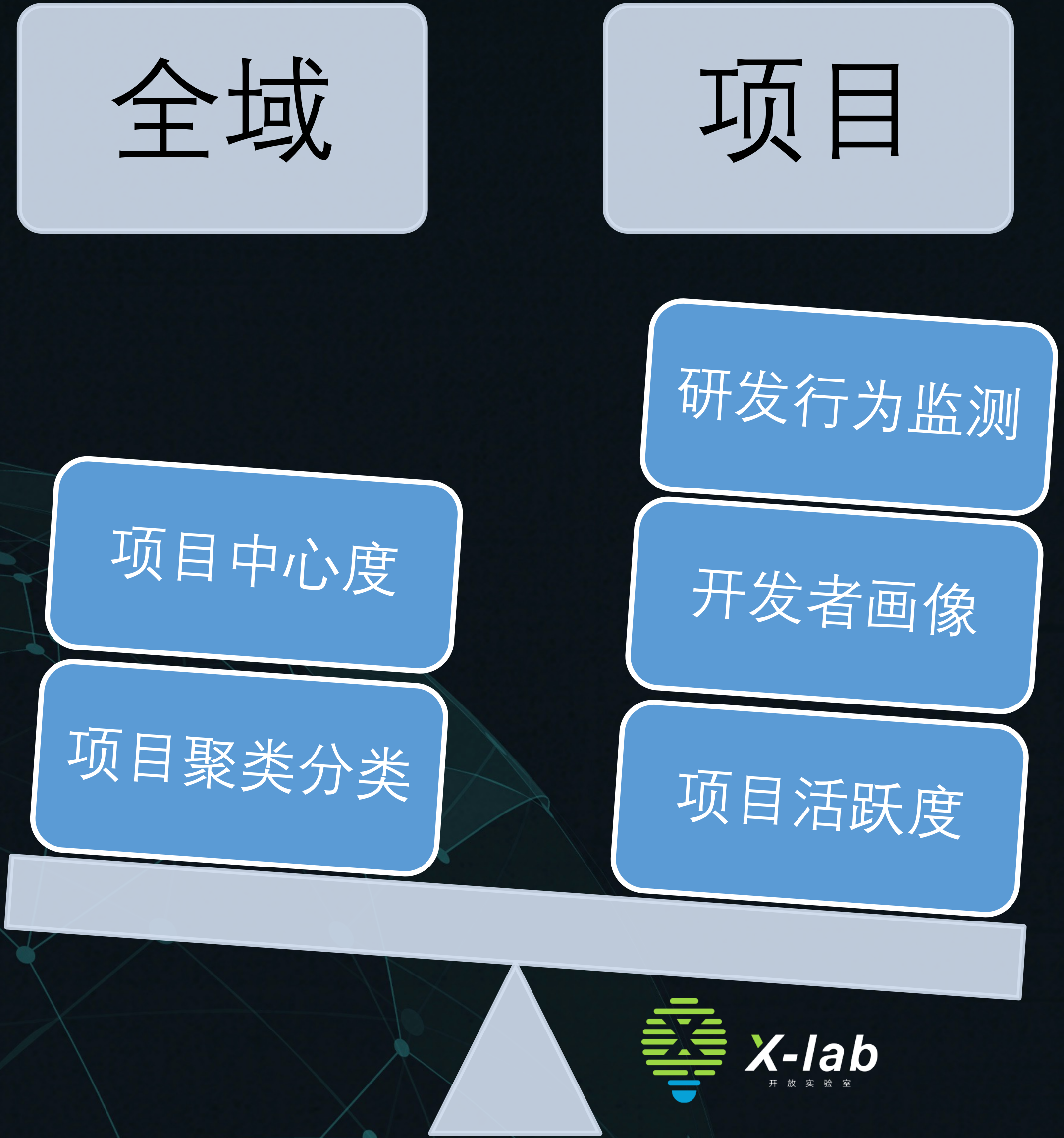
03 问题 Problems

全域

- 项目聚类分类
  - 开源项目分布在哪些领域
  - 给定一个项目，如何预测其所属领域
  - 根据项目类型，可描述开发者画像
- 项目中心度
  - 开源世界中，哪些项目更加重要
  - 项目选型/项目价值预估

项目

- 项目活跃度
  - 项目活跃情况如何？趋势如何？
- 开发者画像
  - 那些参与项目的人都是谁，他们是做什么的？
  - 谁在关注该项目，有什么诉求？
- 研发行为监测
  - 项目是否存在数据异常，有哪些需要注意的风险点？



X-lab 开放实验室

03 问题

Problems

基于 GitHub 的基本操作：开发者活跃于项目 Who-When-What

→ 开发者在项目中活跃，提 Issue，回复 Issue，提 PR，回复 PR，review PR，PR 合入

→ 开发者在具体实例上的活跃度 $A_{di} = \sum w_{ai} C_{ai}, w_{ai} = 2,1,3,1,4,5$??? 权重

→ 开发者在项目上的活跃度 $A_{dr} = \sqrt{\sum A_{di}}$

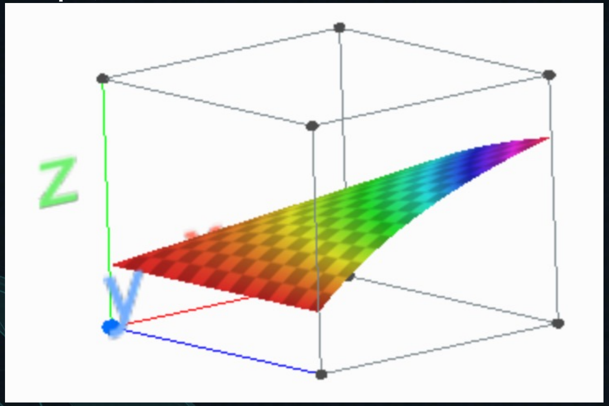
→ 项目的总体活跃度 $A_r = \sum A_{dr}$

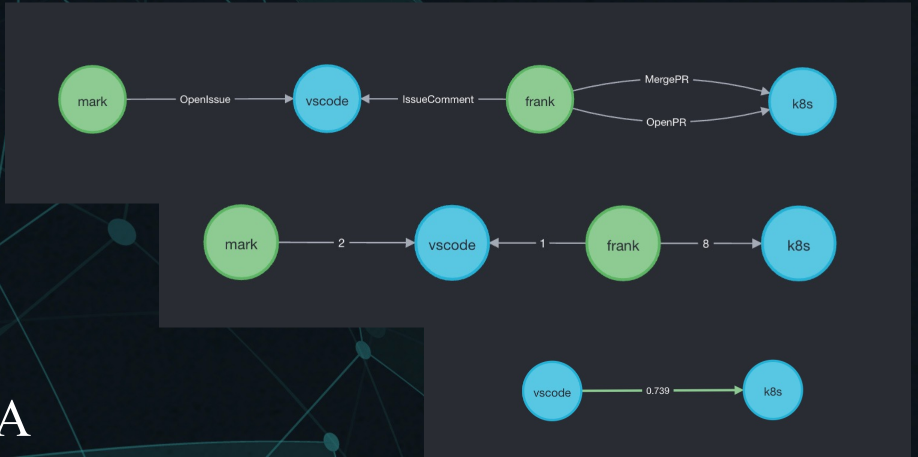
→ 同一项目中开发者之间的协作关联度 $RDP_{ab} = \sum \frac{A_{ai} A_{bi}}{A_{ai} + A_{bi}}$

→ 开发者在全域上的协作关联度 $RDG_{ab} = \sum RDP_{ab}$

→ 不同项目间的协作关联度 $RP_{ab} = \sum \frac{A_a A_b}{A_a + A_b}$??? 聚合

→ 上述均为一段时间内的统计数据，某时间点：180 天 EMA



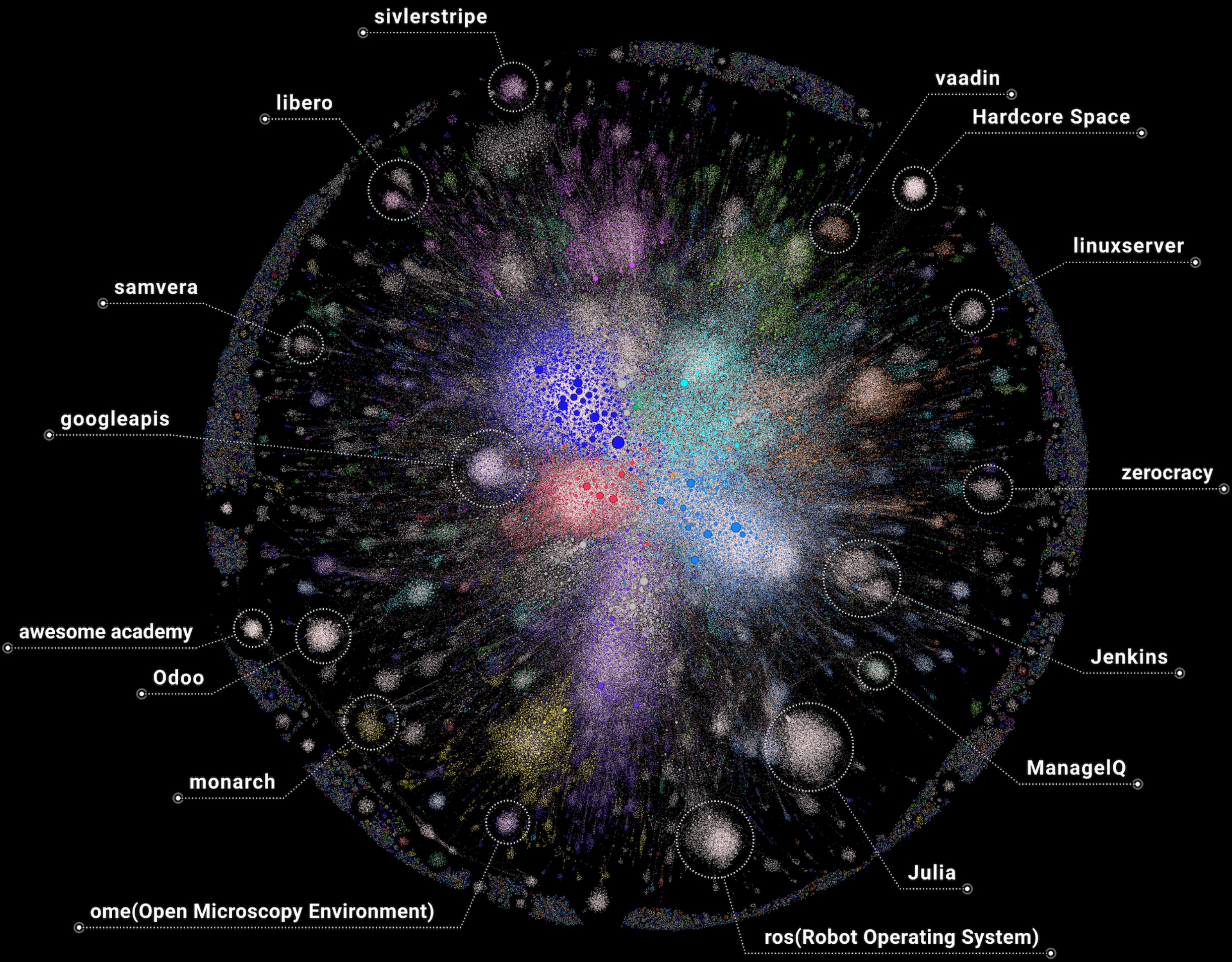


OpenGalaxy 2019

OpenGalaxy is generated by collaboration network of all active GitHub repos in 2019. This graph contains 171,141 nodes and 2,811,489 edges. The generate method can be found in here [1] and the data is from GHArchive [2].

OpenGalaxy 是通过 GitHub 2019 年全域所有活跃项目的协作网络生成的。本图共包含 171,141 个节点和 2,811,489 条边。具体生成方法请参见这里 [1]，数据来自于 GHArchive [2]。

	Area/领域	Top Repos/顶级项目	Count/项目数量
🟦	ts & frontend	VSCode, TypeScript, react, jest	23,254
🟦	cloud native	kubernetes, go, helm, ansible	15,787
🟪	AI libs	pandas, numpy, conda, openjournals	14,971
🟦	tools	rust, nextcloud, godotengine	13,361
🟪	PHP	symfony, laravel, wordpress, magento	8,158
🟥	Microsoft	azure-docs, AspNetCore, WSL	6,276
🟧	system	homebrew, systemd	6,193
🟨	biotech	rstudio, bioconda	6,102
🟩	blockchain	bitcoin, ethereum, ipfs	5,141



[1] http://blog.frankzhao.cn/open_rank_and_open_galaxy/

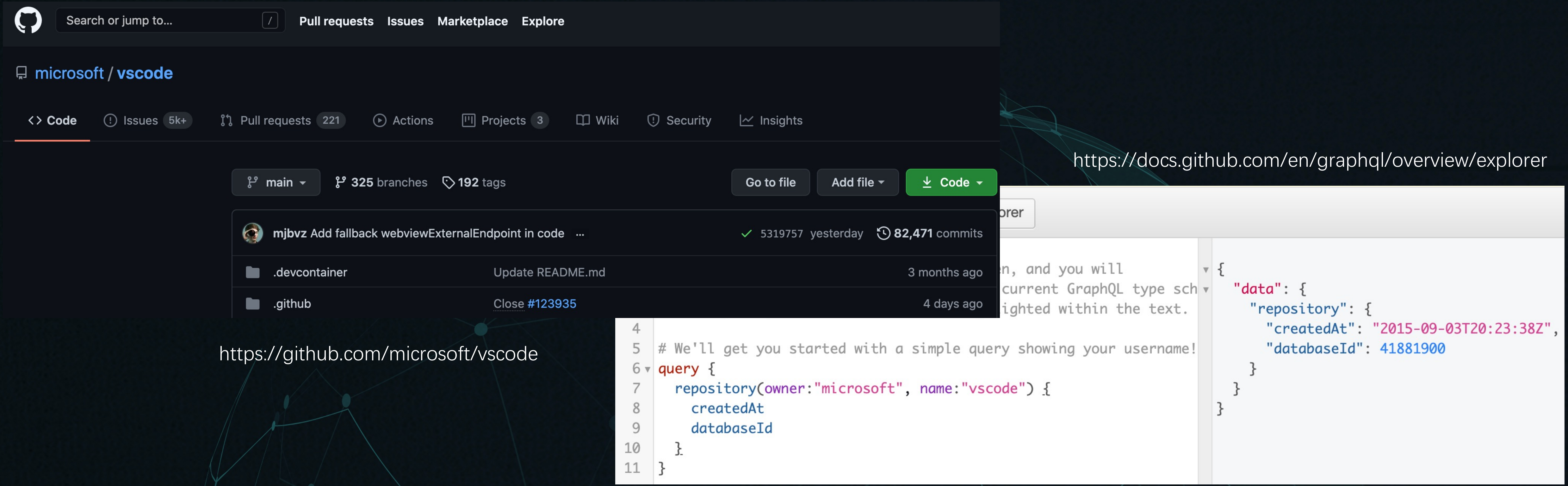
[2] http://www.gharchive.org/

04 举个例子

Example

VSCode

- 开源世界的核心，最流行 IDE
- 准备工作：获取项目的地址、创建时间、数据库 ID



```
n, and you will
current GraphQL type sch
ighted within the text.

# We'll get you started with a simple query showing your username!
query {
  repository(owner:"microsoft", name:"vscode") {
    createdAt
    databaseId
  }
}
```

```
{
  "data": {
    "repository": {
      "createdAt": "2015-09-03T20:23:38Z",
      "databaseId": 41881900
    }
  }
}
```

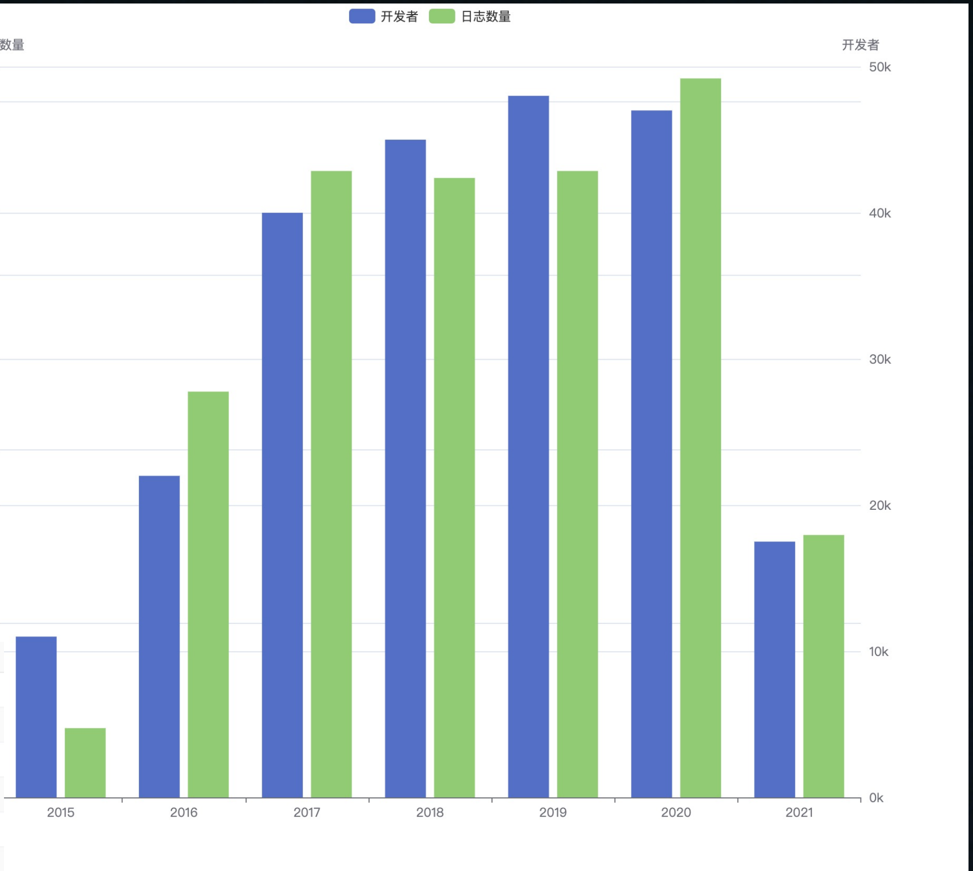
https://github.com/microsoft/vscode

https://docs.github.com/en/graphql/overview/explorer

04 举个例子

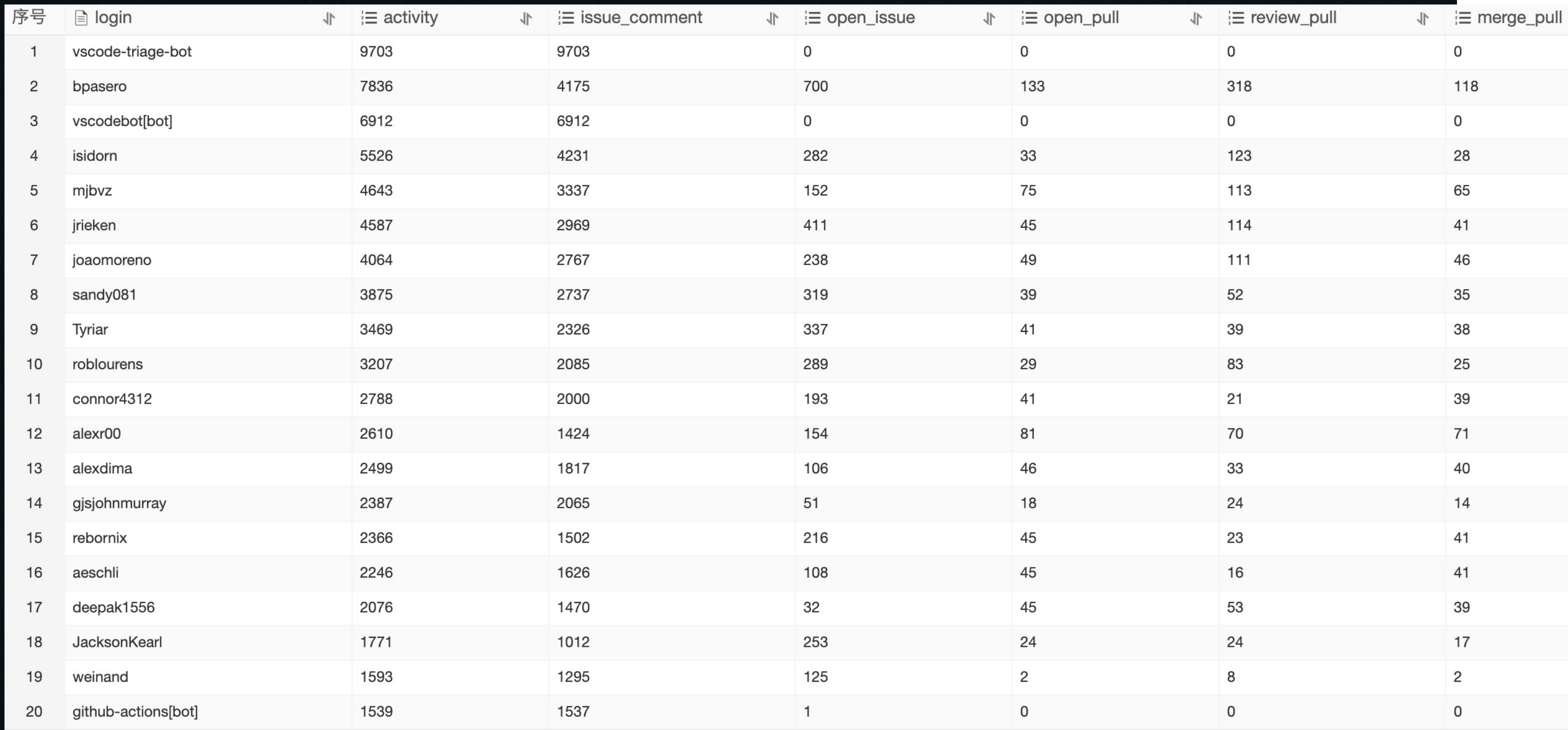
Example

VSCode 案例分析



VSCode 历史日志/开发者数量

VSCode 2020 开发者活跃度



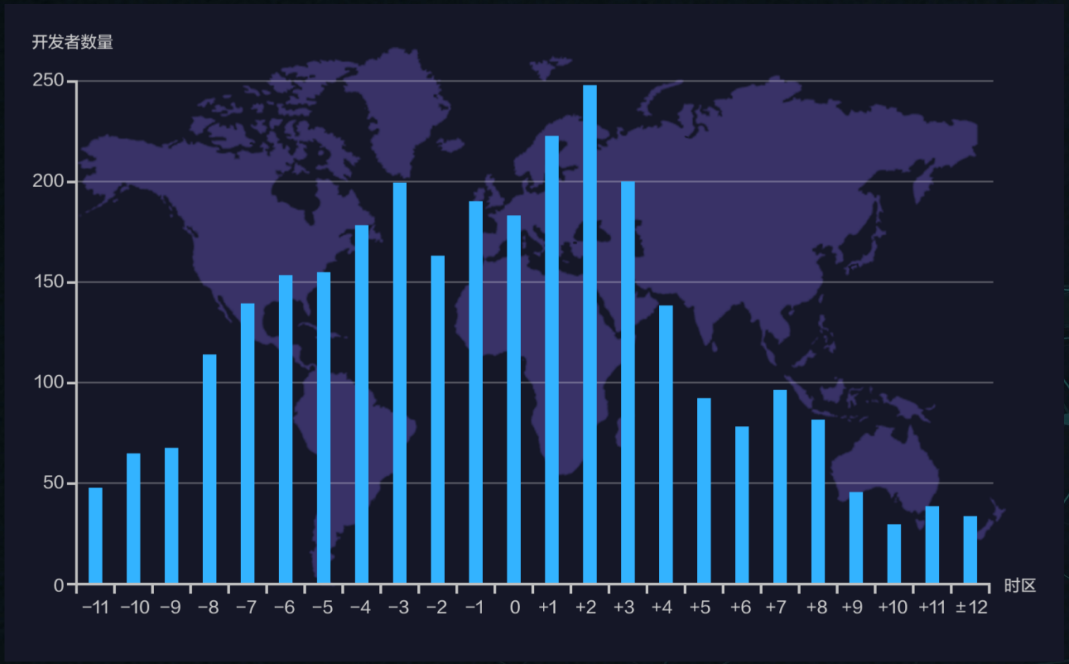
序号	login	activity	issue_comment	open_issue	open_pull	review_pull	merge_pull
1	vscode-triage-bot	9703	9703	0	0	0	0
2	bpasero	7836	4175	700	133	318	118
3	vscodebot[bot]	6912	6912	0	0	0	0
4	isidorn	5526	4231	282	33	123	28
5	mjbvz	4643	3337	152	75	113	65
6	jrieken	4587	2969	411	45	114	41
7	joaomoreno	4064	2767	238	49	111	46
8	sandy081	3875	2737	319	39	52	35
9	Tyriar	3469	2326	337	41	39	38
10	roblourens	3207	2085	289	29	83	25
11	connor4312	2788	2000	193	41	21	39
12	alexr00	2610	1424	154	81	70	71
13	alexdima	2499	1817	106	46	33	40
14	gjsjohnmurray	2387	2065	51	18	24	14
15	rebornix	2366	1502	216	45	23	41
16	aeschli	2246	1626	108	45	16	41
17	deepak1556	2076	1470	32	45	53	39
18	JacksonKearl	1771	1012	253	24	24	17
19	weinand	1593	1295	125	2	8	2
20	github-actions[bot]	1539	1537	1	0	0	0

X-lab 开放实验室

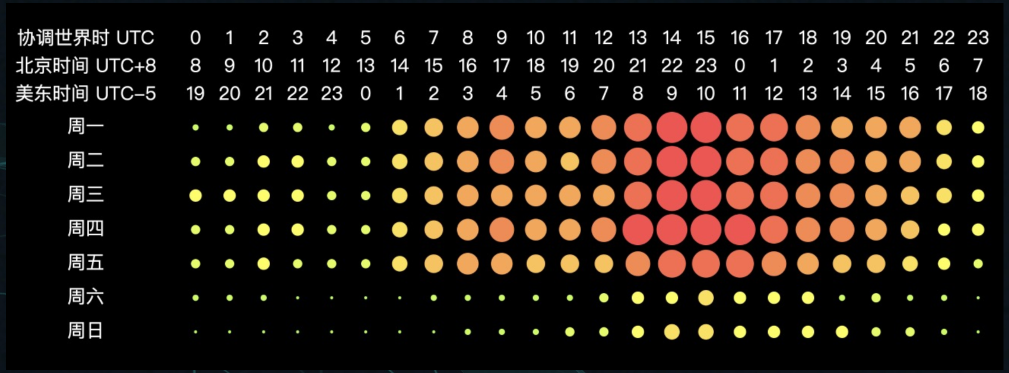
04 / 举个例子

Example

VSCode 案例分析



VSCode 2020 开发者时区分布



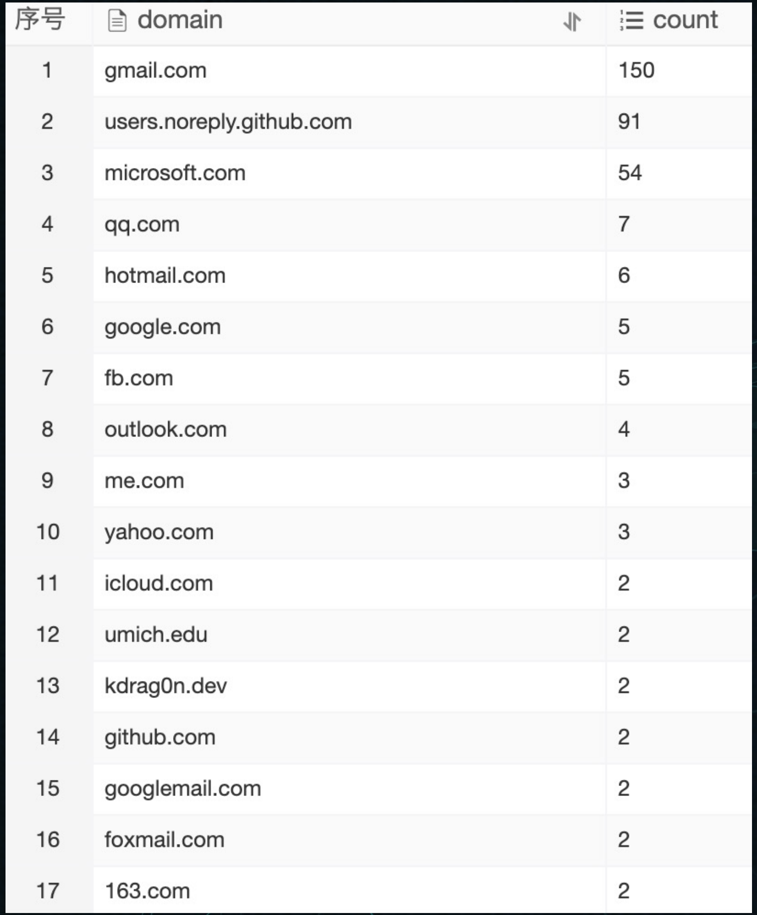
协调世界时 UTC	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
北京时间 UTC+8	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	0	1	2	3	4	5	6	7
美东时间 UTC−5	19	20	21	22	23	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

VSCode 2020 工作时间分布

04 举个例子

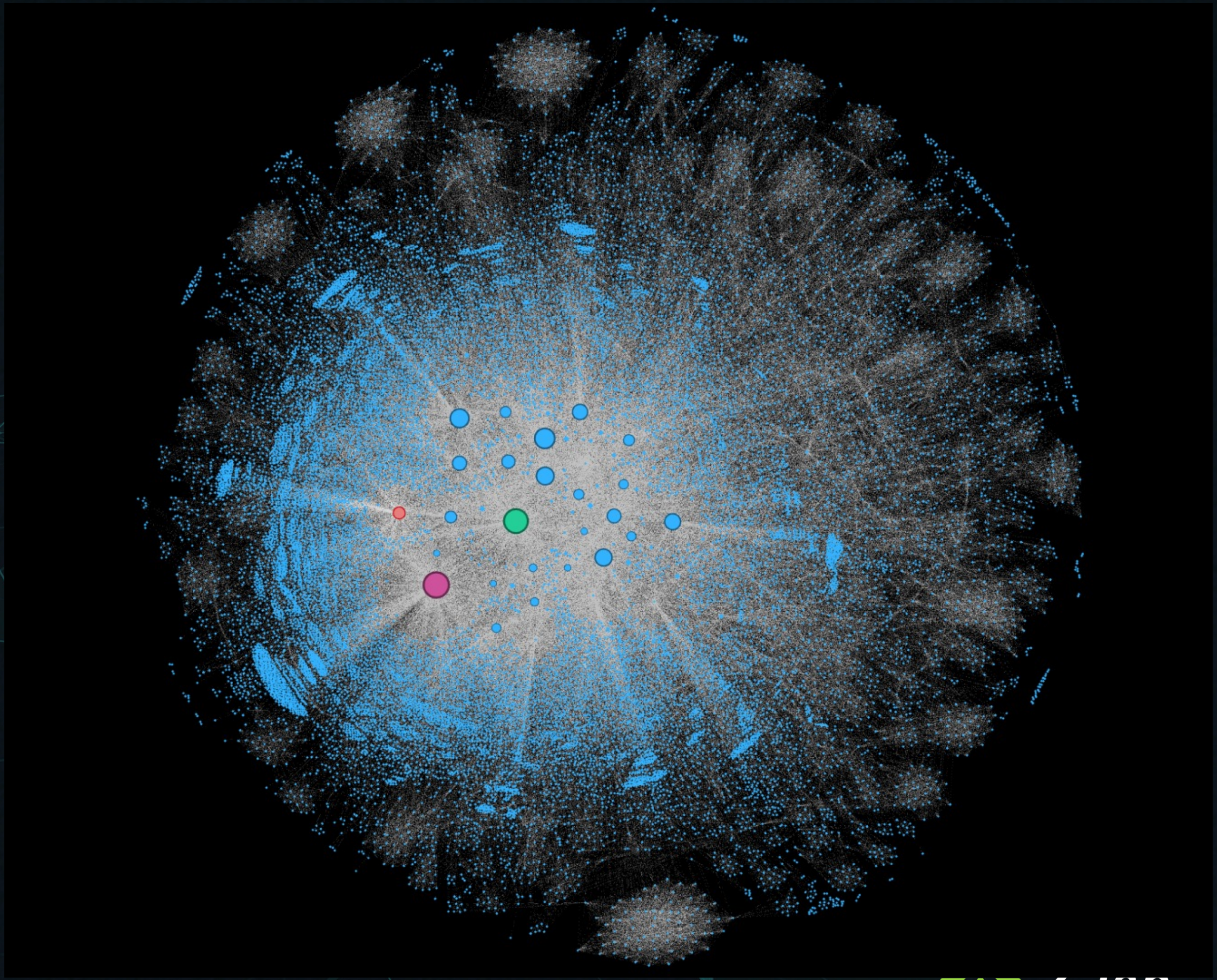
Example

VSCode 案例分析



序号	domain	count
1	gmail.com	150
2	users.noreply.github.com	91
3	microsoft.com	54
4	qq.com	7
5	hotmail.com	6
6	google.com	5
7	fb.com	5
8	outlook.com	4
9	me.com	3
10	yahoo.com	3
11	icloud.com	2
12	umich.edu	2
13	kdrag0n.dev	2
14	github.com	2
15	googlemail.com	2
16	foxmail.com	2
17	163.com	2

VSCode 2020 开发者邮箱后缀分布

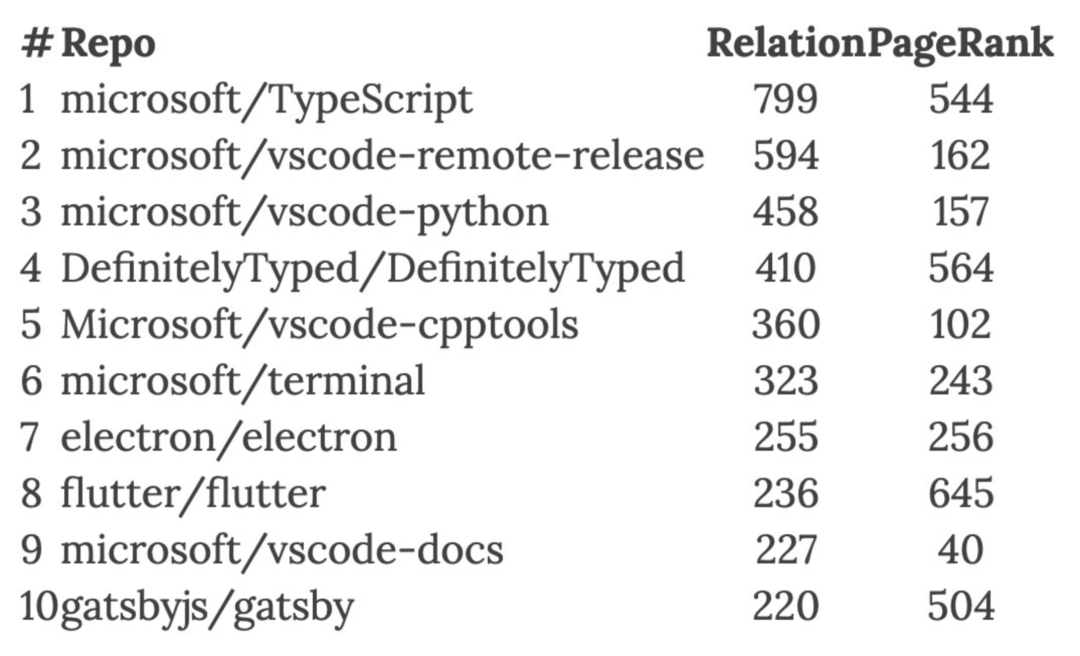


VSCode 2020 开发者协作网络

04 举个例子

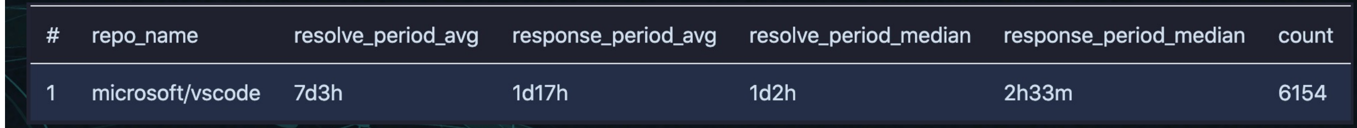
Example

VSCode 案例分析



#	Repo	Relation	PageRank
1	microsoft/TypeScript	799	544
2	microsoft/vscode-remote-release	594	162
3	microsoft/vscode-python	458	157
4	DefinitelyTyped/DefinitelyTyped	410	564
5	Microsoft/vscode-cpptools	360	102
6	microsoft/terminal	323	243
7	electron/electron	255	256
8	flutter/flutter	236	645
9	microsoft/vscode-docs	227	40
10	gatsbyjs/gatsby	220	504

VSCode 2020 项目协作关联 Top10



#	repo_name	resolve_period_avg	response_period_avg	resolve_period_median	response_period_median	count
1	microsoft/vscode	7d3h	1d17h	1d2h	2h33m	6154

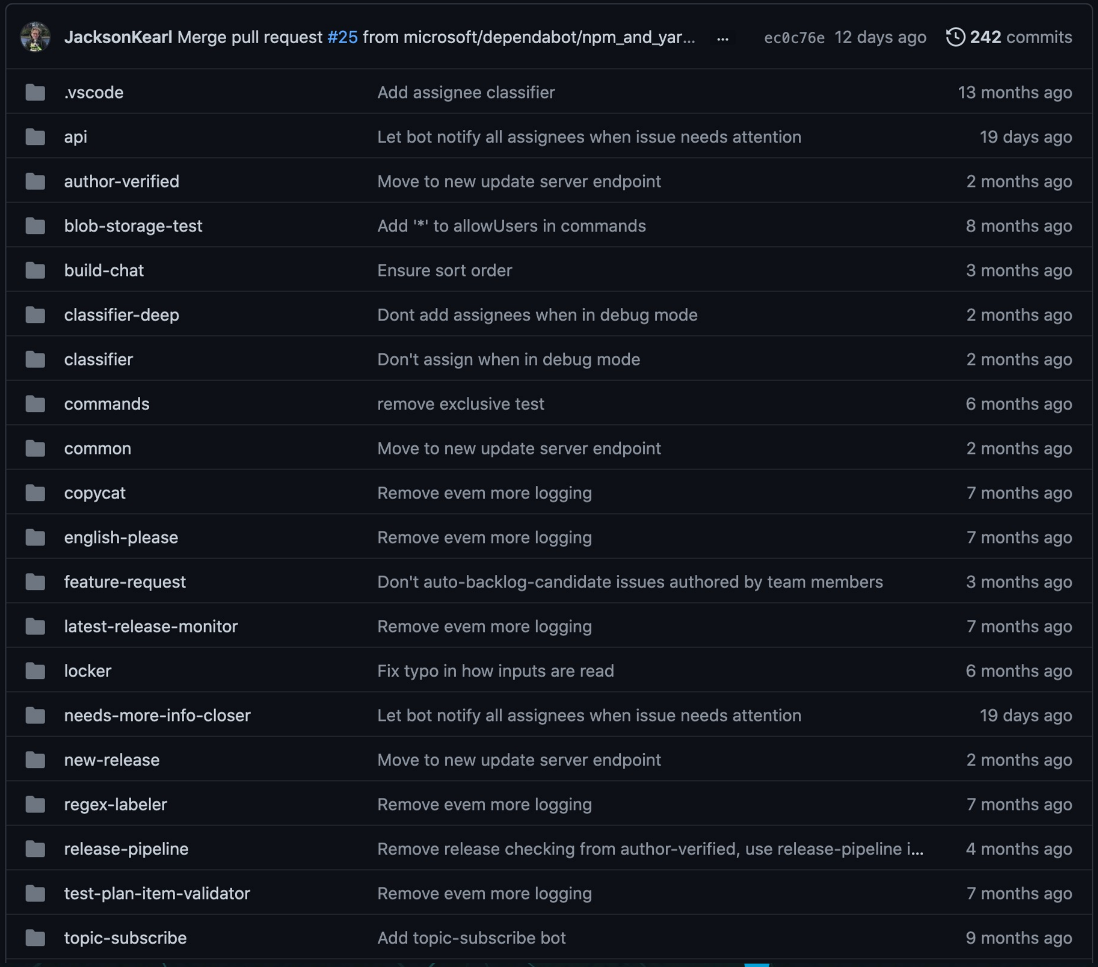
VSCode 2021 Issue 响应解决周期

04 / 举个例子

Example

VSCode 案例分析之社区流程

- build-chat：将构建信息发送到 Slack 中
- classifier/classifier-deep：Issue 自动打标/基于机器学习
- copycat：跨仓库 Issue 拷贝
- english-please：非英文开 Issue 提示使用英文
- locker：Issue 关闭一段时间后自动锁定
- needs-more-info-closer：需要用户反馈的 Issue 若一段时间没有回复自动关闭
- regex-labeler：根据 Issue 描述中正则匹配结果打标签
- topic-subscribe：根据 label 提醒某些账户关注当前 Issue



JacksonKearl Merge pull request #25 from microsoft/dependabot/npm_and_yar... ··· ec0c76e 12 days ago 242 commits

.vscode	Add assignee classifier	13 months ago
api	Let bot notify all assignees when issue needs attention	19 days ago
author-verified	Move to new update server endpoint	2 months ago
blob-storage-test	Add '*' to allowUsers in commands	8 months ago
build-chat	Ensure sort order	3 months ago
classifier-deep	Dont add assignees when in debug mode	2 months ago
classifier	Don't assign when in debug mode	2 months ago
commands	remove exclusive test	6 months ago
common	Move to new update server endpoint	2 months ago
copycat	Remove evem more logging	7 months ago
english-please	Remove evem more logging	7 months ago
feature-request	Don't auto-backlog-candidate issues authored by team members	3 months ago
latest-release-monitor	Remove evem more logging	7 months ago
locker	Fix typo in how inputs are read	6 months ago
needs-more-info-closer	Let bot notify all assignees when issue needs attention	19 days ago
new-release	Move to new update server endpoint	2 months ago
regex-labeler	Remove evem more logging	7 months ago
release-pipeline	Remove release checking from author-verified, use release-pipeline i...	4 months ago
test-plan-item-validator	Remove evem more logging	7 months ago
topic-subscribe	Add topic-subscribe bot	9 months ago

https://github.com/microsoft/vscode-github-triage-actions/

另一些注重流程的有趣社区：Kubernetes、React、OpenDigger

05 大作业

Assignment

某开源项目的 2020 年深入数据分析

数据类

- 基础的统计数据分析、可视化
- 开发者数据统计、可视化
- 关联数据的分析，如协作关联度高的其他项目
- 其他任意想做的数据分析

流程类

- 项目的日常协作流程调研
- 开发者参与流程调研
- 项目 CI/CD 的流程调研

课程提供

- 全域数据的 Clickhouse 数据库只读访问能力
- 预置常用数据统计的 SQL
- 分析过程中的指导答疑

https://github.com/X-lab2017/open-digger

THANK YOU

X-lab
开放实验室