Beyond bus-factor

data-visualizations and algorithms to better understand Apache communities

What’s going on?

What’s going on?

(in the community)

Márton Elek

- Apache Ozone /Hadoop/Ratis PMC
  - elek@apache.org
- twitter.com/@anzix
- Ozone:
  - Code and Console (Youtube)
  - Ozone Explained (Youtube)
- Kubernetes + Apache Bigdata:
  - **github.com/elek**/flekszible
  - flokkr.github.io

S3 protocol

Hadoop FS

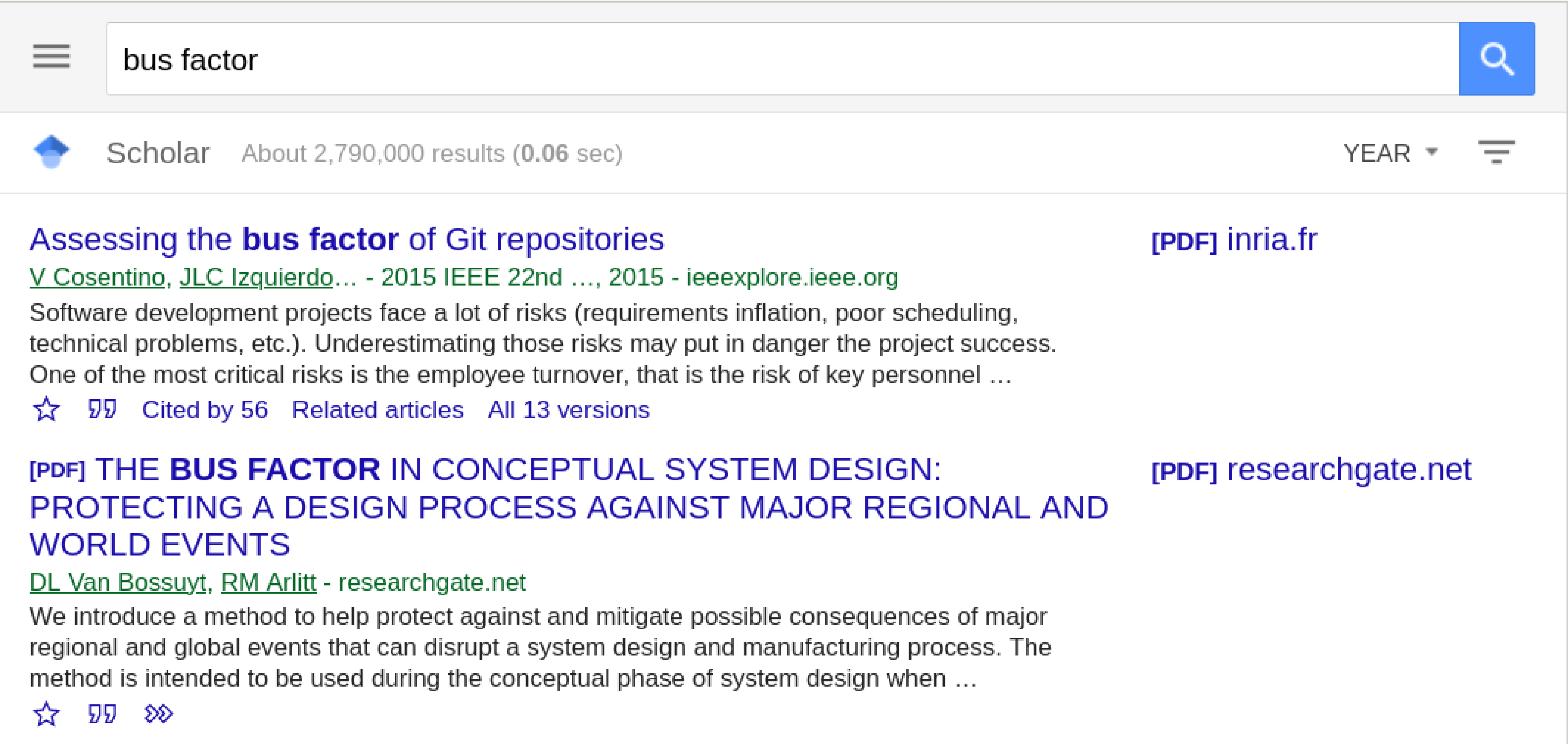
CSI

Apache Hadoop Ozone

ozone.apache.org

The "bus factor" is the **minimum number** of team members that have to suddenly disappear from a project **before** the **project stalls** due to lack of knowledgeable or competent personnel.

(Wikipedia)



About 2,790,000 results (**0.06** sec)

Assessing the **bus factor** of Git repositories — [PDF] inria.fr

V Cosentino, JLC Izquierdo… - 2015 IEEE 22nd …, 2015 - ieeexplore.ieee.org

Software development projects face a lot of risks (requirements inflation, poor scheduling, technical problems, etc.). Underestimating those risks may put in danger the project success. One of the most critical risks is the employee turnover, that is the risk of key personnel …

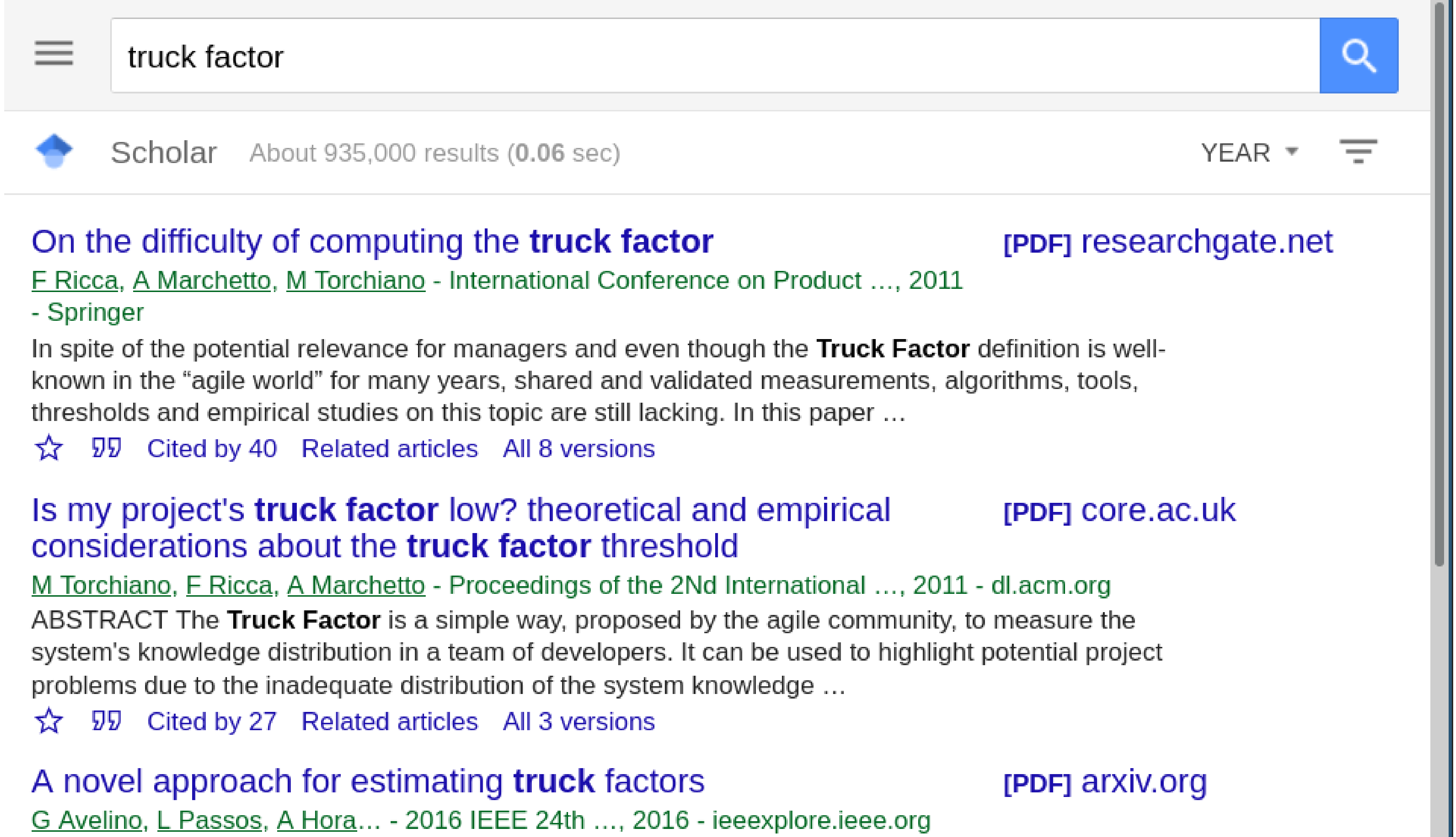
☆ Cited by 56 Related articles All 13 versions

[PDF] THE **BUS FACTOR** IN CONCEPTUAL SYSTEM DESIGN: PROTECTING A DESIGN PROCESS AGAINST MAJOR REGIONAL AND WORLD EVENTS — [PDF] researchgate.net

DL Van Bossuyt, RM Arlitt - researchgate.net

We introduce a method to help protect against and mitigate possible consequences of major regional and global events that can disrupt a system design and manufacturing process. The method is intended to be used during the conceptual phase of system design when …

☆



truck factor

About 935,000 results (**0.06** sec)

On the difficulty of computing the **truck factor** [PDF] researchgate.net

F Ricca, A Marchetto, M Torchiano - International Conference on Product …, 2011 - Springer

In spite of the potential relevance for managers and even though the **Truck Factor** definition is well-known in the "agile world" for many years, shared and validated measurements, algorithms, tools, thresholds and empirical studies on this topic are still lacking. In this paper …

☆ ❞ Cited by 40 Related articles All 8 versions

Is my project's **truck factor** low? theoretical and empirical considerations about the **truck factor** threshold [PDF] core.ac.uk

M Torchiano, F Ricca, A Marchetto - Proceedings of the 2Nd International …, 2011 - dl.acm.org

ABSTRACT The **Truck Factor** is a simple way, proposed by the agile community, to measure the system's knowledge distribution in a team of developers. It can be used to highlight potential project problems due to the inadequate distribution of the system knowledge …

☆ ❞ Cited by 27 Related articles All 3 versions

A novel approach for estimating **truck** factors [PDF] arxiv.org

G Avelino, L Passos, A Hora… - 2016 IEEE 24th …, 2016 - ieeexplore.ieee.org

~~Bus~~ Leave factor

A Novel Approach for Estimating Truck Factors

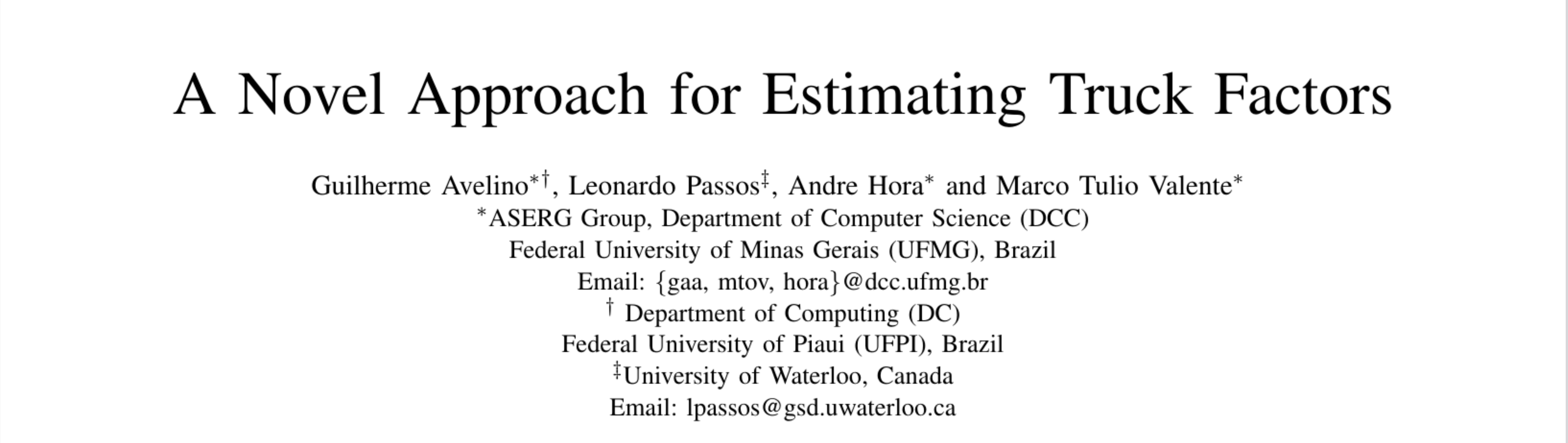
Guilherme Avelino[*][†], Leonardo Passos[‡], Andre Hora[*] and Marco Tulio Valente[*]
[*]ASERG Group, Department of Computer Science (DCC)
Federal University of Minas Gerais (UFMG), Brazil
Email: {gaa, mtov, hora}@dcc.ufmg.br
[†] Department of Computing (DC)
Federal University of Piaui (UFPI), Brazil
[‡]University of Waterloo, Canada
Email: lpassos@gsd.uwaterloo.ca

***Abstract*—Truck Factor (TF) is a metric proposed by the agile community as a tool to identify concentration of knowledge in software development environments. It states the minimal number of developers that have to be hit by a truck (or quit) before a project is incapacitated. In other words, TF helps to measure how prepared is a project to deal with developer turnover. Despite its clear relevance, few studies explore this metric. Altogether there is no consensus about how to calculate it, and no supporting evidence backing estimates for systems in the wild. To mitigate both issues, we propose a novel (and automated) approach for estimating TF-values, which we execute against a corpus of 133 popular project in GitHub. We later survey developers as a means to assess the reliability of our results.**

for TF-estimation for which we apply to a target corpus comprising 133 systems in GitHub. In total, such systems have over 373K files and 41 MLOC; their combined evolution history sums to over 2 million commits. By surveying and analyzing answers from 67 target systems, we evidence that in 84% of valid answers developers agree or partially agree that the TF's authors are the main authors of their systems; in 53% we receive a positive or partially positive answer regarding our estimated truck factors.

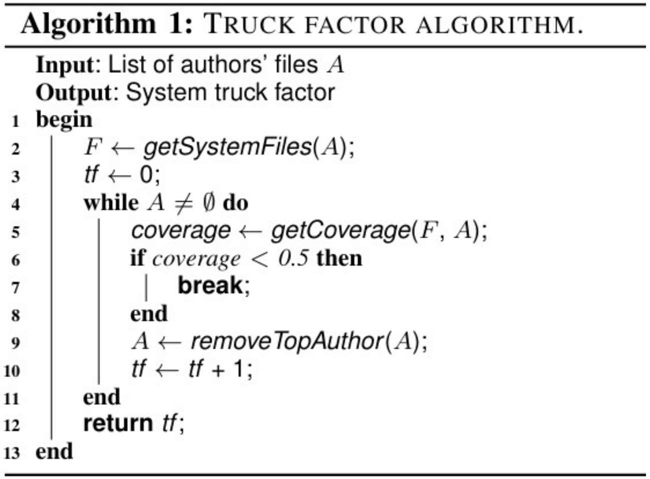
From our work, we claim the following contributions:

1) A novel approach for estimating a system's truck factor



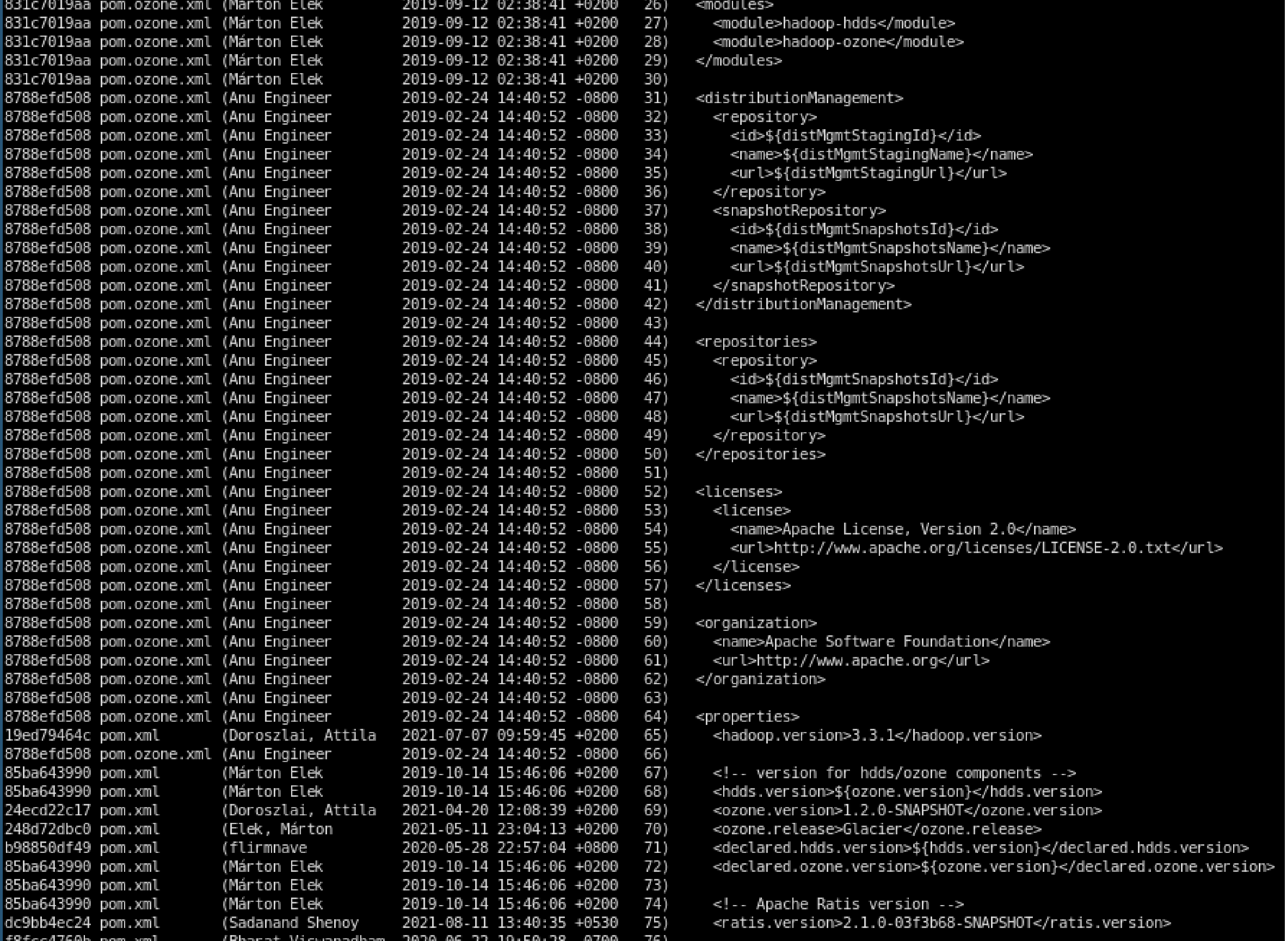
A Novel Approach for Estimating Truck Factors

Guilherme Avelino*†, Leonardo Passos‡, Andre Hora* and Marco Tulio Valente*
*ASERG Group, Department of Computer Science (DCC)
Federal University of Minas Gerais (UFMG), Brazil
Email: {gaa, mtov, hora}@dcc.ufmg.br
† Department of Computing (DC)
Federal University of Piaui (UFPI), Brazil
‡University of Waterloo, Canada
Email: lpassos@gsd.uwaterloo.ca



Algorithm 1: TRUCK FACTOR ALGORITHM.

```
Input: List of authors' files $A$
Output: System truck factor
begin
    $F \leftarrow$ getSystemFiles($A$);
    $tf \leftarrow 0$;
    while $A \neq \emptyset$ do
        coverage $\leftarrow$ getCoverage($F$, $A$);
        if $coverage < 0.5$ then
            break;
        end
        $A \leftarrow$ removeTopAuthor($A$);
        $tf \leftarrow tf + 1$;
    end
    return $tf$;
end
```



```
831c7019aa pom.ozone.xml (Márton Elek         2019-09-12 02:38:41 +0200   26)   <modules>
831c7019aa pom.ozone.xml (Márton Elek         2019-09-12 02:38:41 +0200   27)     <module>hadoop-hdds</module>
831c7019aa pom.ozone.xml (Márton Elek         2019-09-12 02:38:41 +0200   28)     <module>hadoop-ozone</module>
831c7019aa pom.ozone.xml (Márton Elek         2019-09-12 02:38:41 +0200   29)   </modules>
831c7019aa pom.ozone.xml (Márton Elek         2019-09-12 02:38:41 +0200   30)
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   31)   <distributionManagement>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   32)     <repository>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   33)       <id>${distMgmtStagingId}</id>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   34)       <name>${distMgmtStagingName}</name>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   35)       <url>${distMgmtStagingUrl}</url>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   36)     </repository>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   37)     <snapshotRepository>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   38)       <id>${distMgmtSnapshotsId}</id>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   39)       <name>${distMgmtSnapshotsName}</name>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   40)       <url>${distMgmtSnapshotsUrl}</url>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   41)     </snapshotRepository>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   42)   </distributionManagement>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   43)
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   44)   <repositories>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   45)     <repository>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   46)       <id>${distMgmtSnapshotsId}</id>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   47)       <name>${distMgmtSnapshotsName}</name>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   48)       <url>${distMgmtSnapshotsUrl}</url>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   49)     </repository>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   50)   </repositories>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   51)
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   52)   <licenses>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   53)     <license>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   54)       <name>Apache License, Version 2.0</name>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   55)       <url>http://www.apache.org/licenses/LICENSE-2.0.txt</url>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   56)     </license>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   57)   </licenses>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   58)
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   59)   <organization>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   60)     <name>Apache Software Foundation</name>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   61)     <url>http://www.apache.org</url>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   62)   </organization>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   63)
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   64)   <properties>
19ed79464c pom.xml       (Doroszlai, Attila   2021-07-07 09:59:45 +0200   65)     <hadoop.version>3.3.1</hadoop.version>
8788efd508 pom.ozone.xml (Anu Engineer        2019-02-24 14:40:52 -0800   66)
85ba643990 pom.xml       (Márton Elek         2019-10-14 15:46:06 +0200   67)     <!-- version for hdds/ozone components -->
85ba643990 pom.xml       (Márton Elek         2019-10-14 15:46:06 +0200   68)     <hdds.version>${ozone.version}</hdds.version>
24ecd22c17 pom.xml       (Doroszlai, Attila   2021-04-20 12:08:39 +0200   69)     <ozone.version>1.2.0-SNAPSHOT</ozone.version>
248d72dbc0 pom.xml       (Elek, Márton        2021-05-11 23:04:13 +0200   70)     <ozone.release>Glacier</ozone.release>
b98850df49 pom.xml       (flirmnave           2020-05-28 22:57:04 +0800   71)     <declared.hdds.version>${hdds.version}</declared.hdds.version>
85ba643990 pom.xml       (Márton Elek         2019-10-14 15:46:06 +0200   72)     <declared.ozone.version>${ozone.version}</declared.ozone.version>
85ba643990 pom.xml       (Márton Elek         2019-10-14 15:46:06 +0200   73)
85ba643990 pom.xml       (Márton Elek         2019-10-14 15:46:06 +0200   74)     <!-- Apache Ratis version -->
dc9bb4ec24 pom.xml       (Sadanand Shenoy     2021-08-11 13:40:35 +0530   75)     <ratis.version>2.1.0-03f3b68-SNAPSHOT</ratis.version>
f8fcc4760b pom.xml       (Bharat Viswanadham  2020-06-22 19:50:28 -0700   76)
```

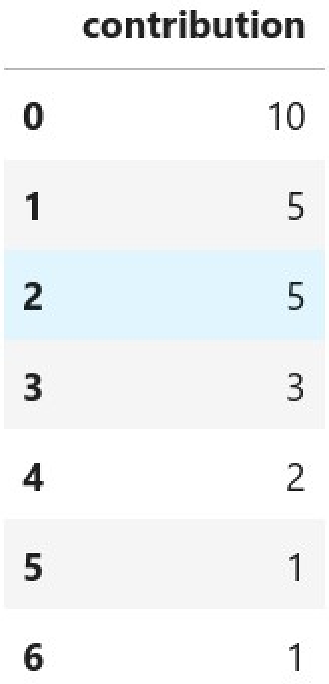
Leave factor / Pony number

- Measurable
- Grouping
- Threshold
- **Number of groups where Σ > threshold**

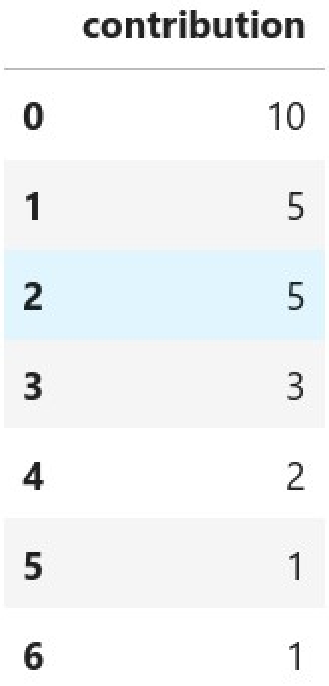
Source: IKEA recipe book

Pony number

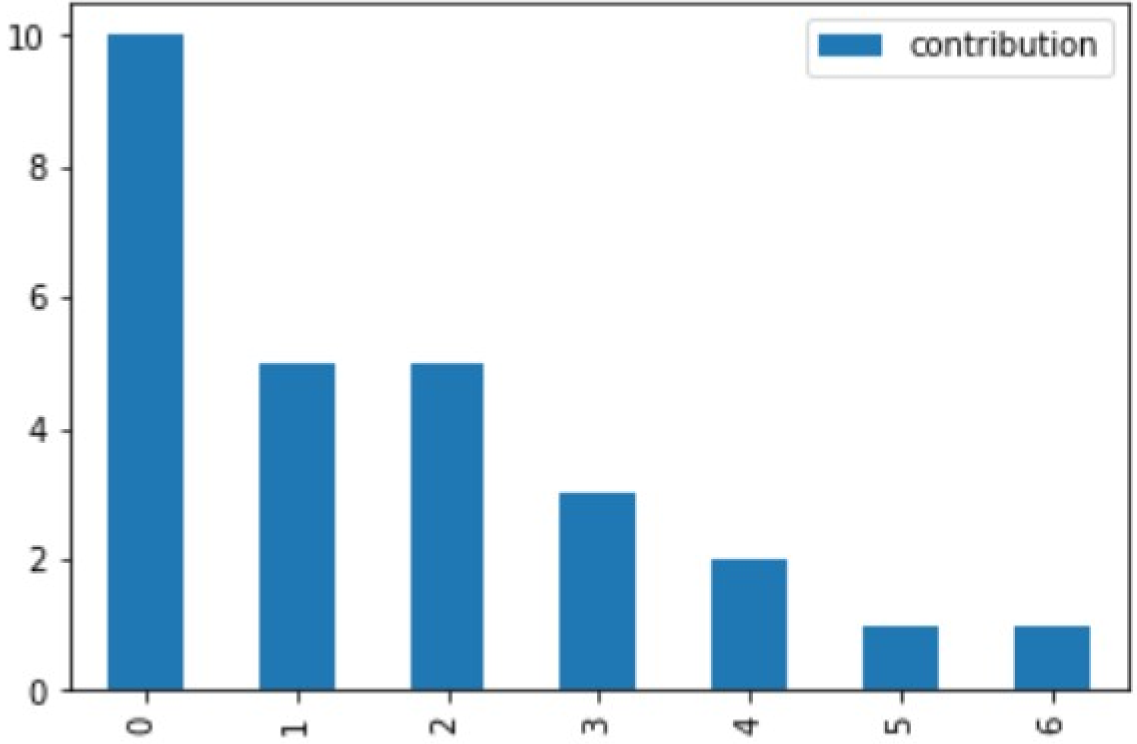
- Measurable: **volume of ingredients**
- Grouping: **ingredient types**
- Threshold: **50%**
- Number of groups where Σ > threshold

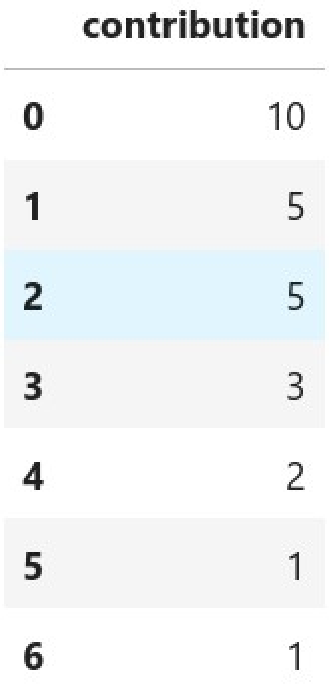


	contribution
0	10
1	5
2	5
3	3
4	2
5	1
6	1

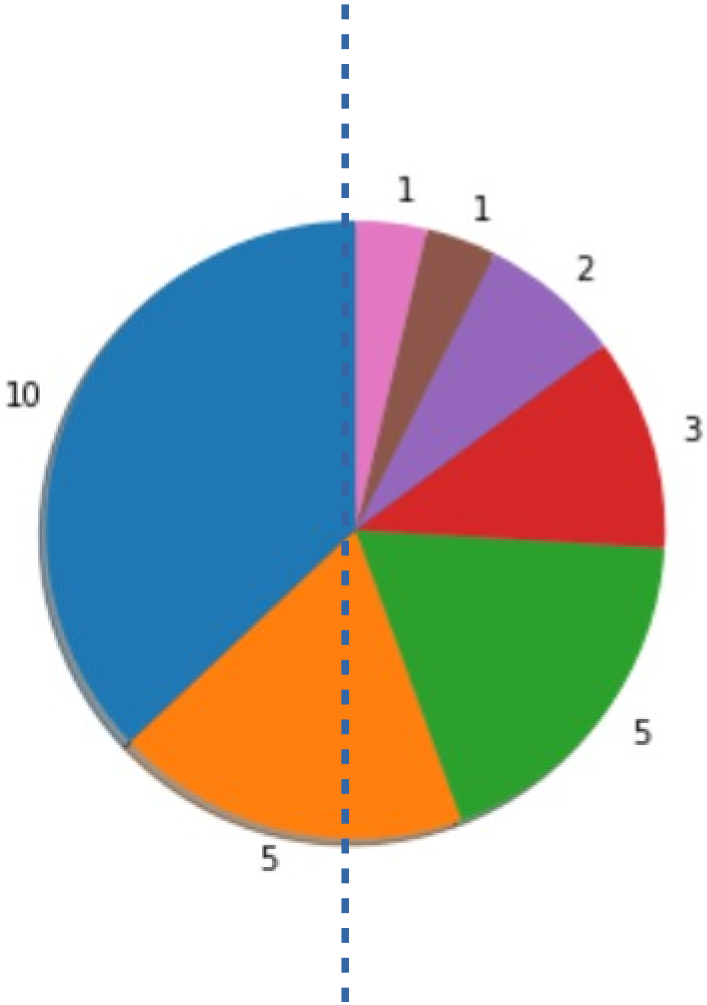


	contribution
0	10
1	5
2	5
3	3
4	2
5	1
6	1



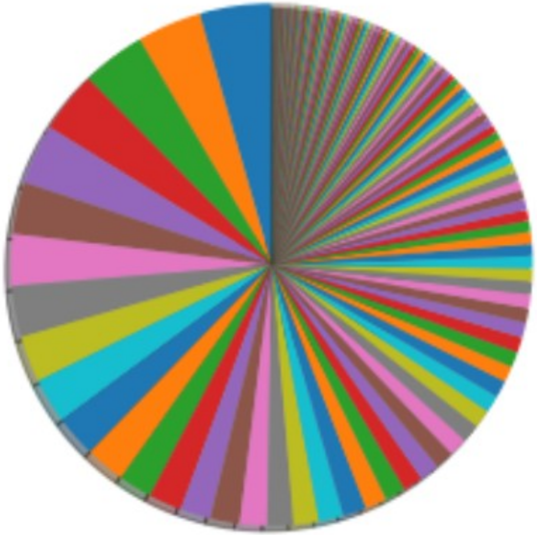


	contribution
0	10
1	5
2	5
3	3
4	2
5	1
6	1

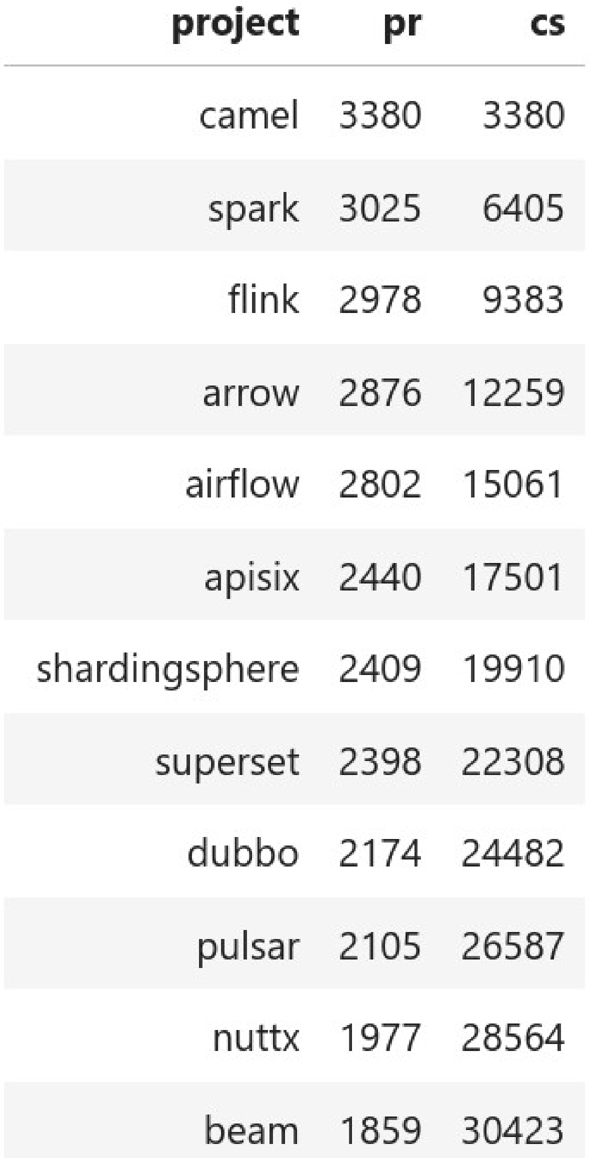


All Apache projects,

prs in this year grouped by projects



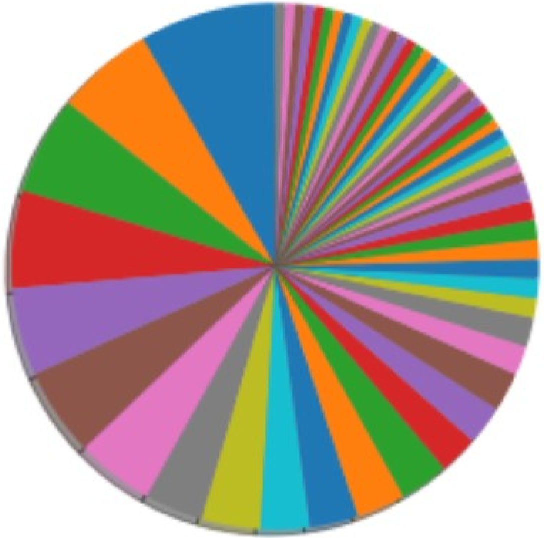
Pony number: 18



project	pr	cs
camel	3380	3380
spark	3025	6405
flink	2978	9383
arrow	2876	12259
airflow	2802	15061
apisix	2440	17501
shardingsphere	2409	19910
superset	2398	22308
dubbo	2174	24482
pulsar	2105	26587
nuttx	1977	28564
beam	1859	30423

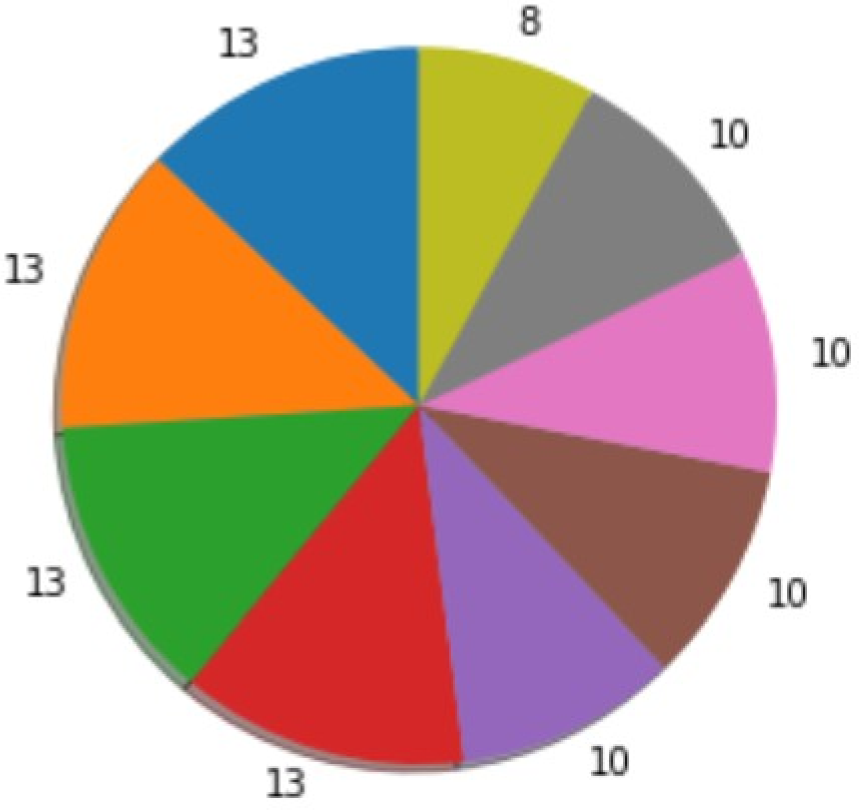
Apache Spark git commits

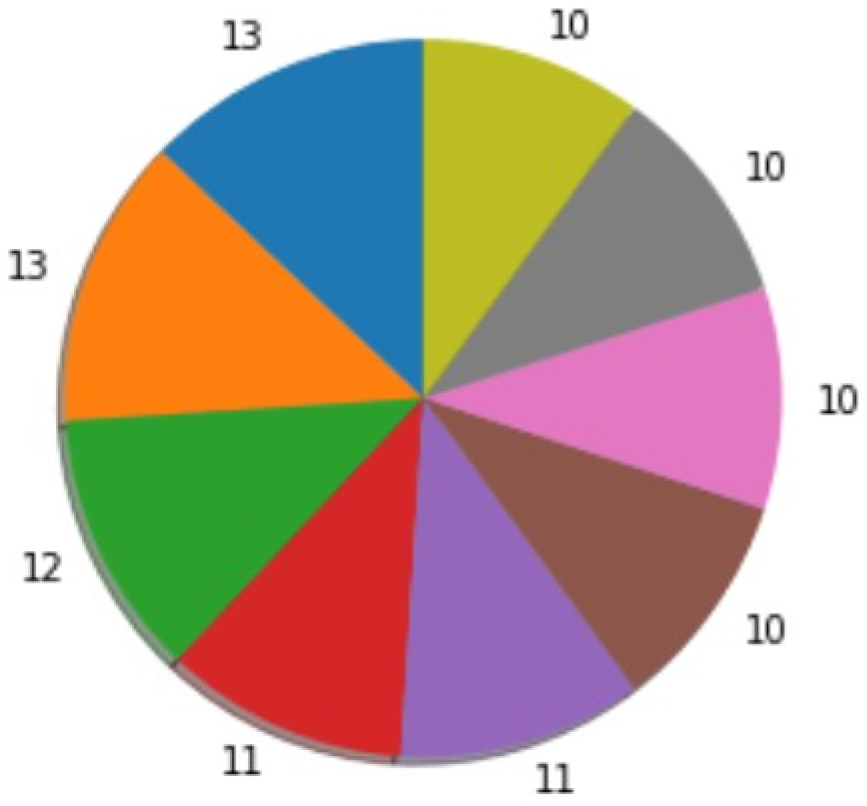
by authors, last 30 days, apache/spark.git



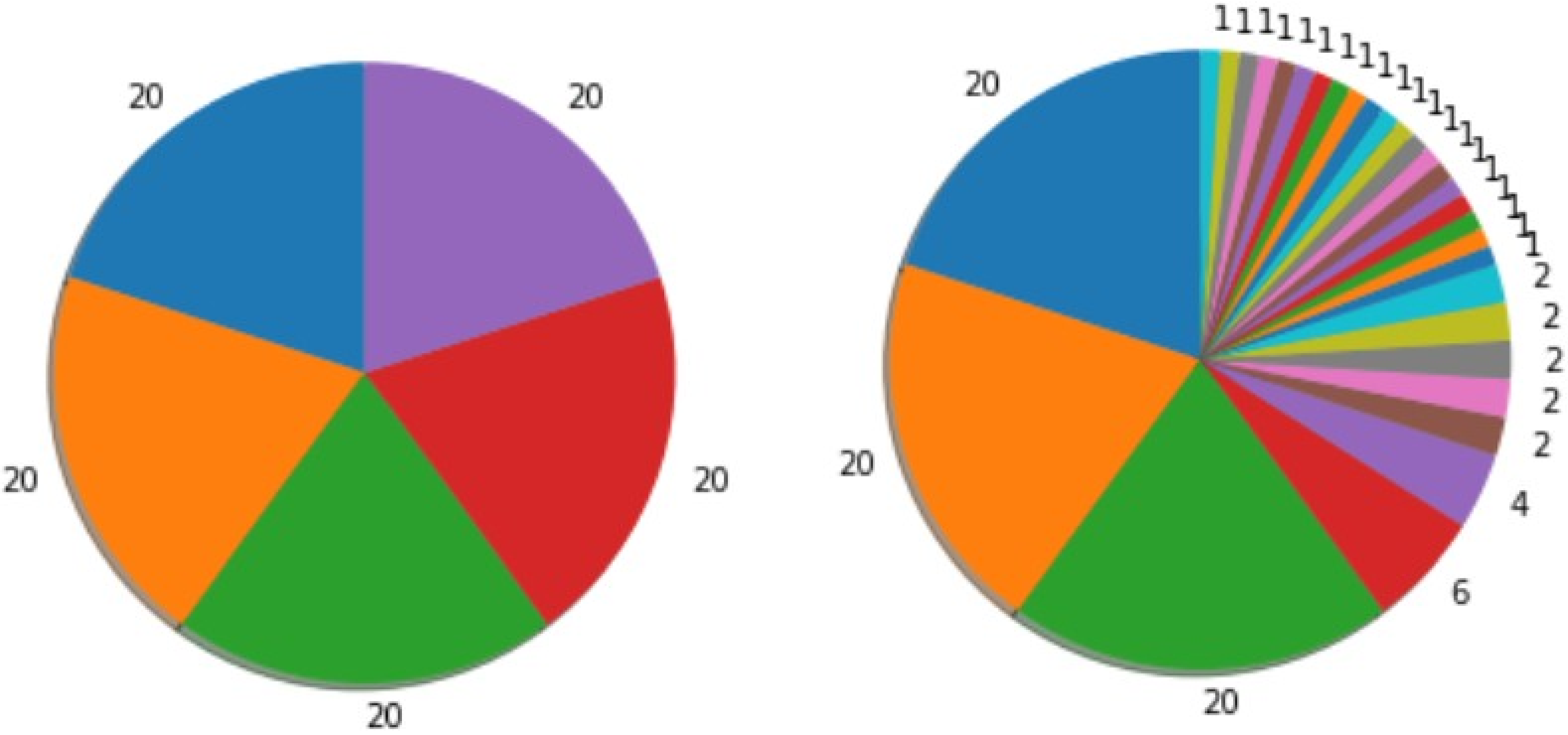
Pony number: 11

Problem #1: Granularity

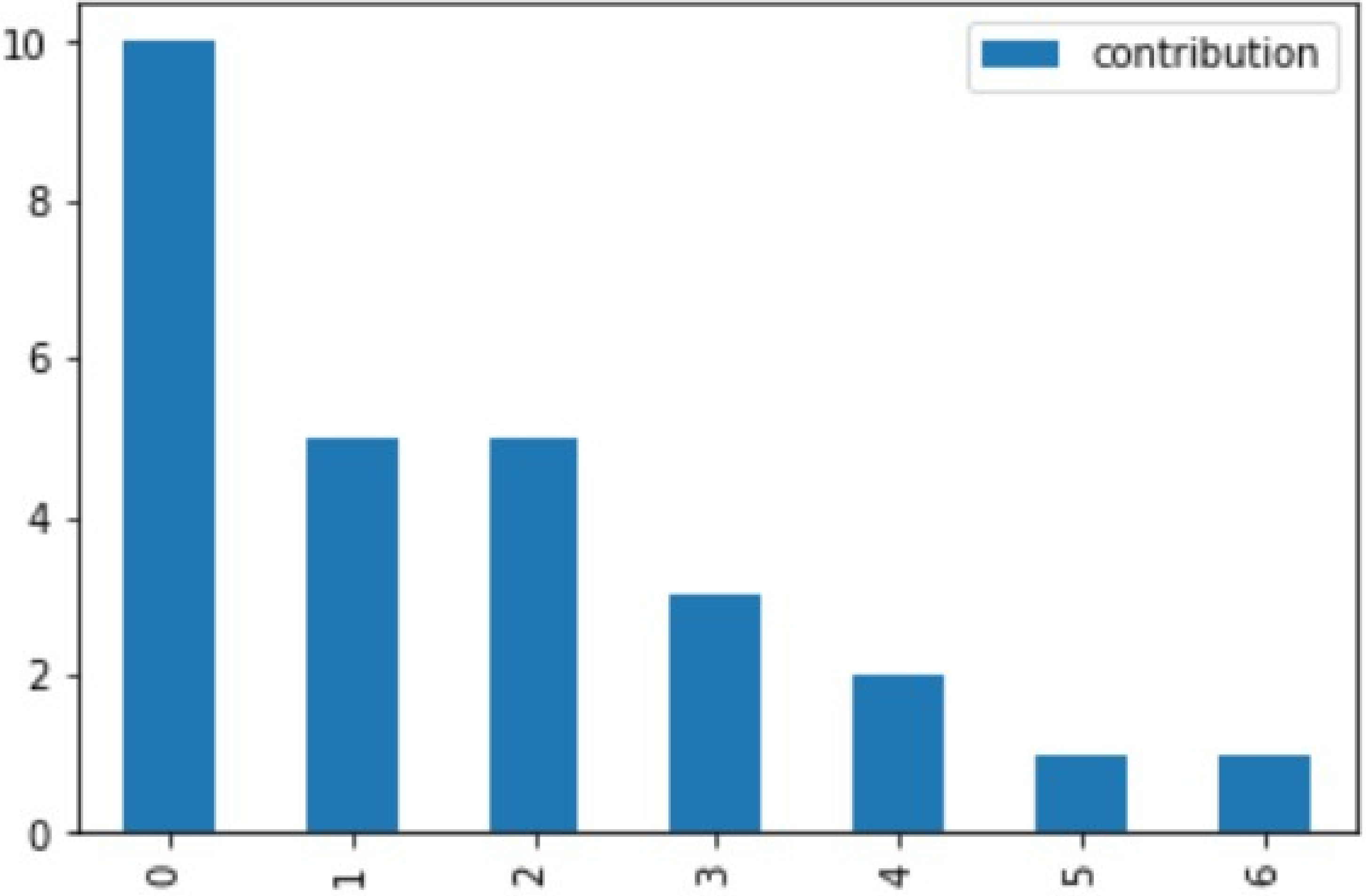


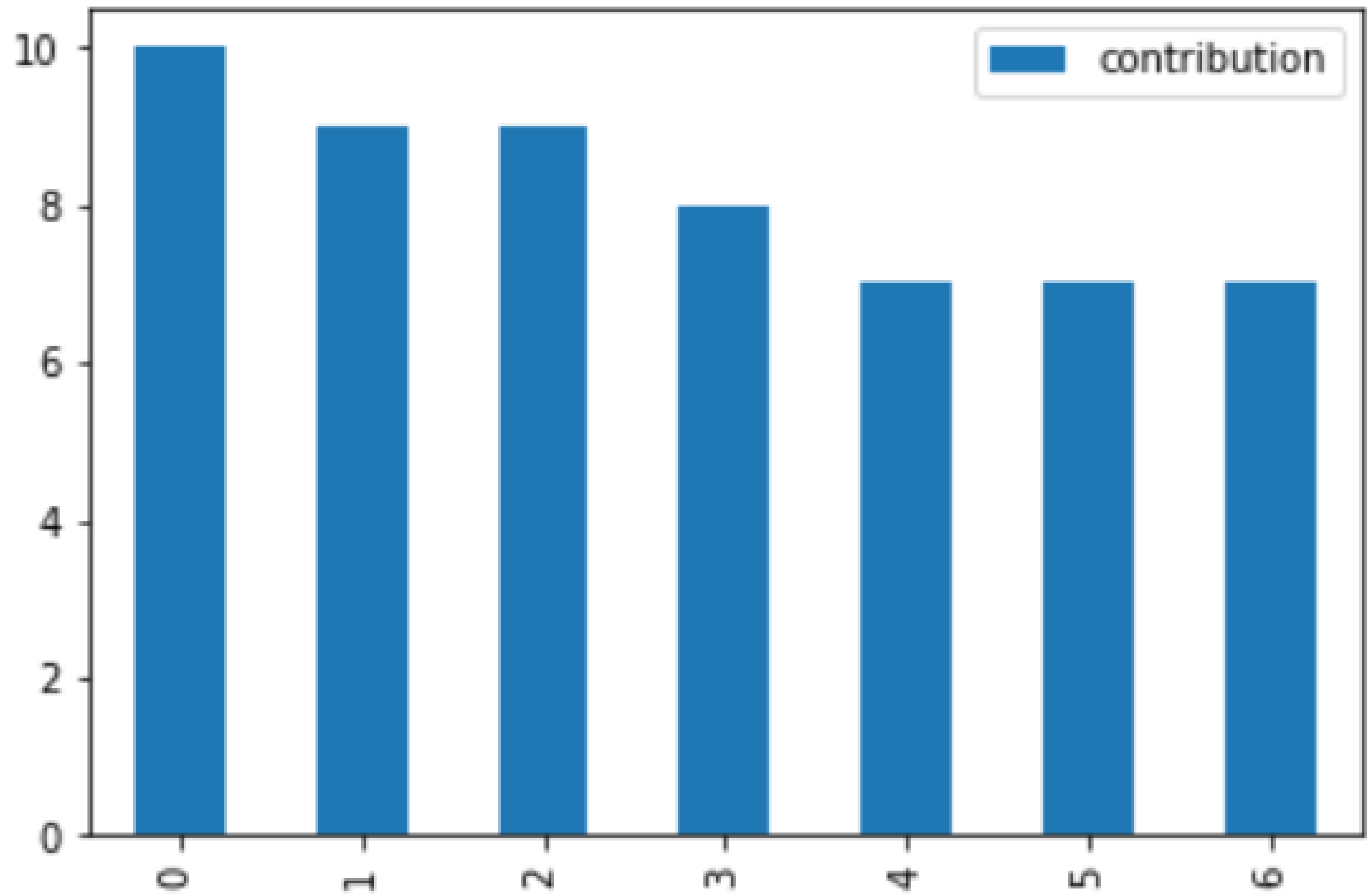


Problem #2: Long tail



Alternatives

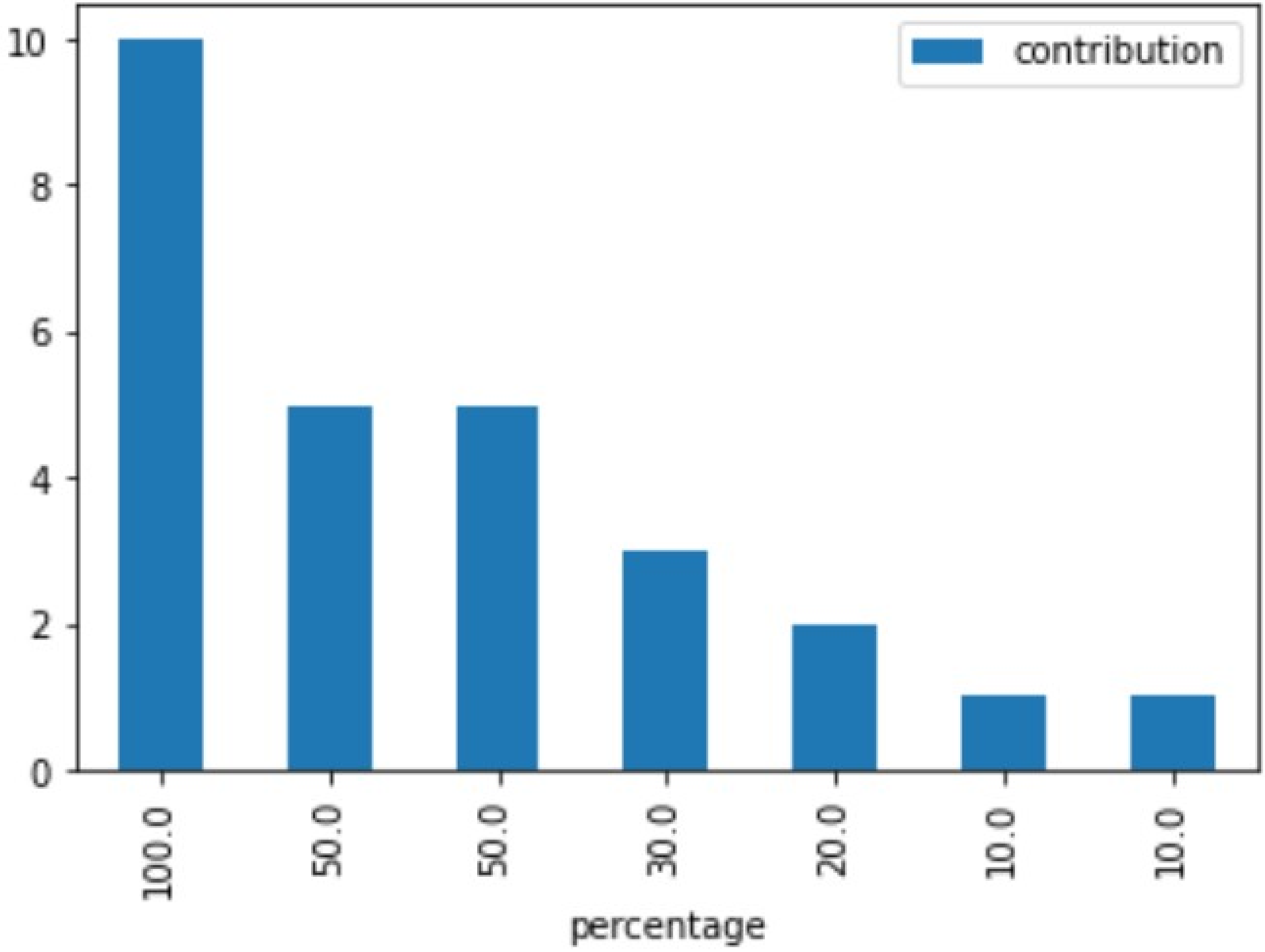

contribution
0
2
4
6
8
10
0
1
2
3
4
5
6

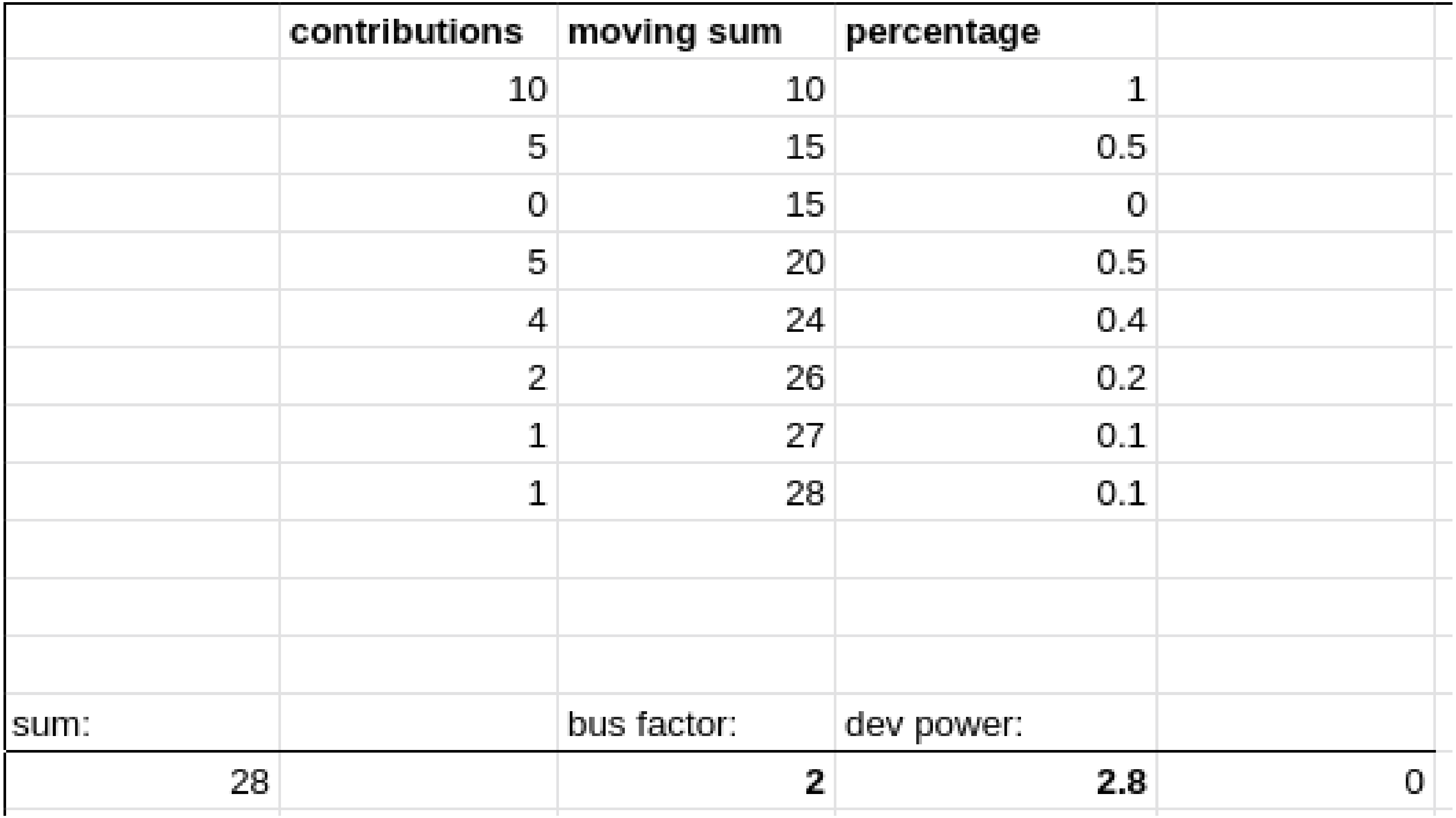


contribution
10
8
6
4
2
0
0
1
2
3
4
5
6

IDEA

- How big is the contribution compared to the highest one?

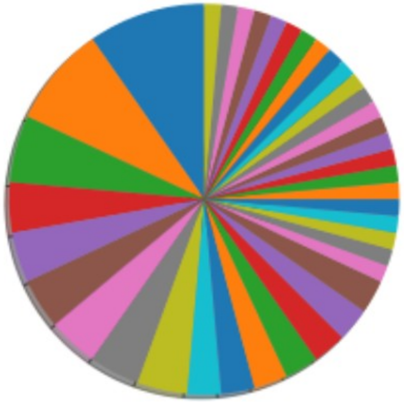




	contributions	moving sum	percentage	
	10	10	1	
	5	15	0.5	
	0	15	0	
	5	20	0.5	
	4	24	0.4	
	2	26	0.2	
	1	27	0.1	
	1	28	0.1	
sum:		bus factor:	dev power:	
28		**2**	**2.8**	0

Git commits by authors, last 30 days

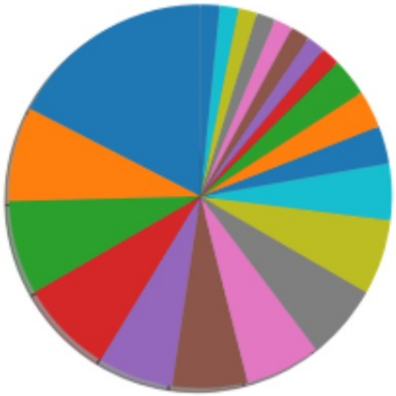
apache/kafka.git



Pony number: 10

First dev ratio: 10.29

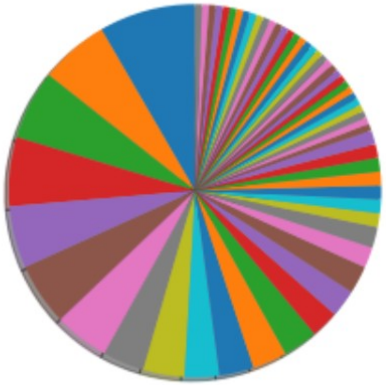
apache/ozone.git



Pony number: 6

First dev ratio: 5.73

apache/spark.git

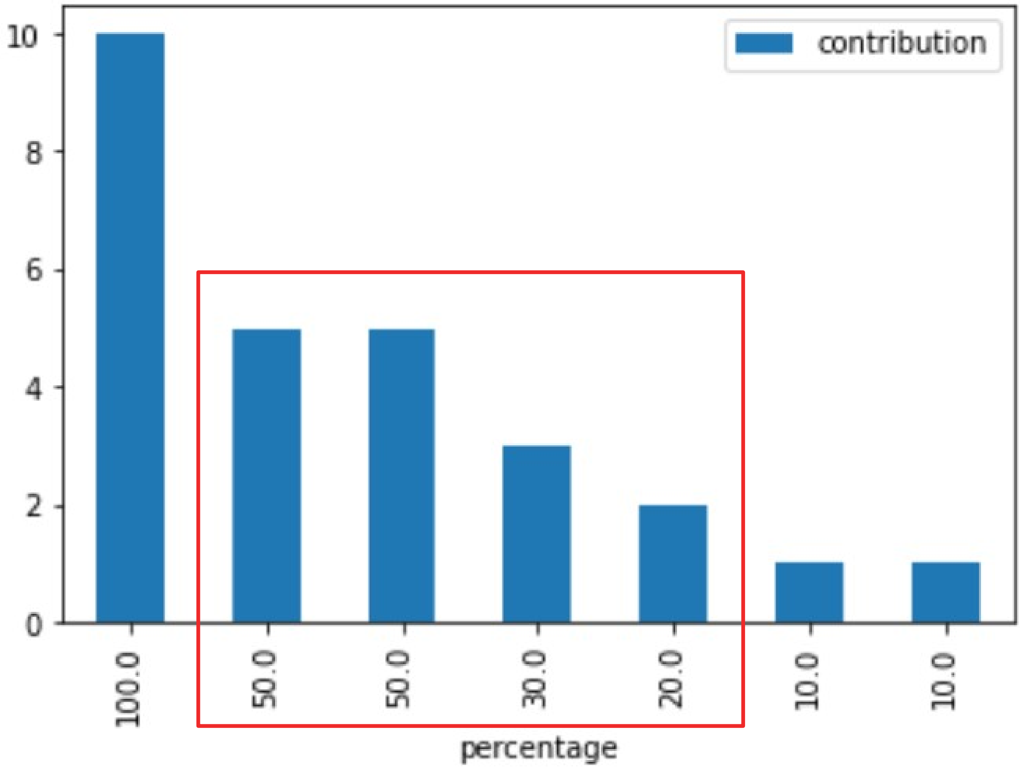


Pony number: 10

First dev ratio: 11.79

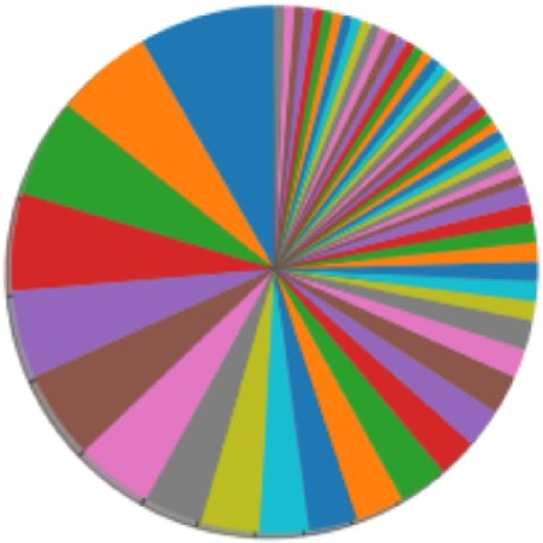
Problems?

- sum(...) / first
- How to evalute the critical mass?



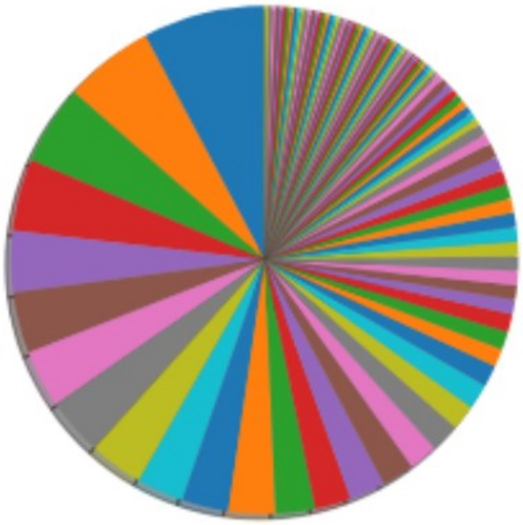
Notes about time

Spark commits last 30/60/180 days



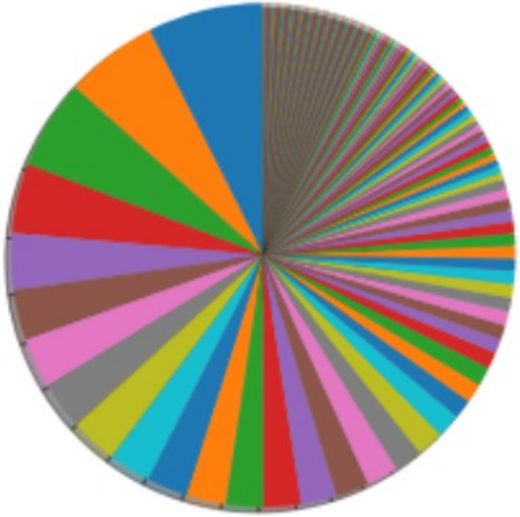
Pony number: 11

First dev ratio: 11.93



Pony number: 12

First dev ratio: 12.82

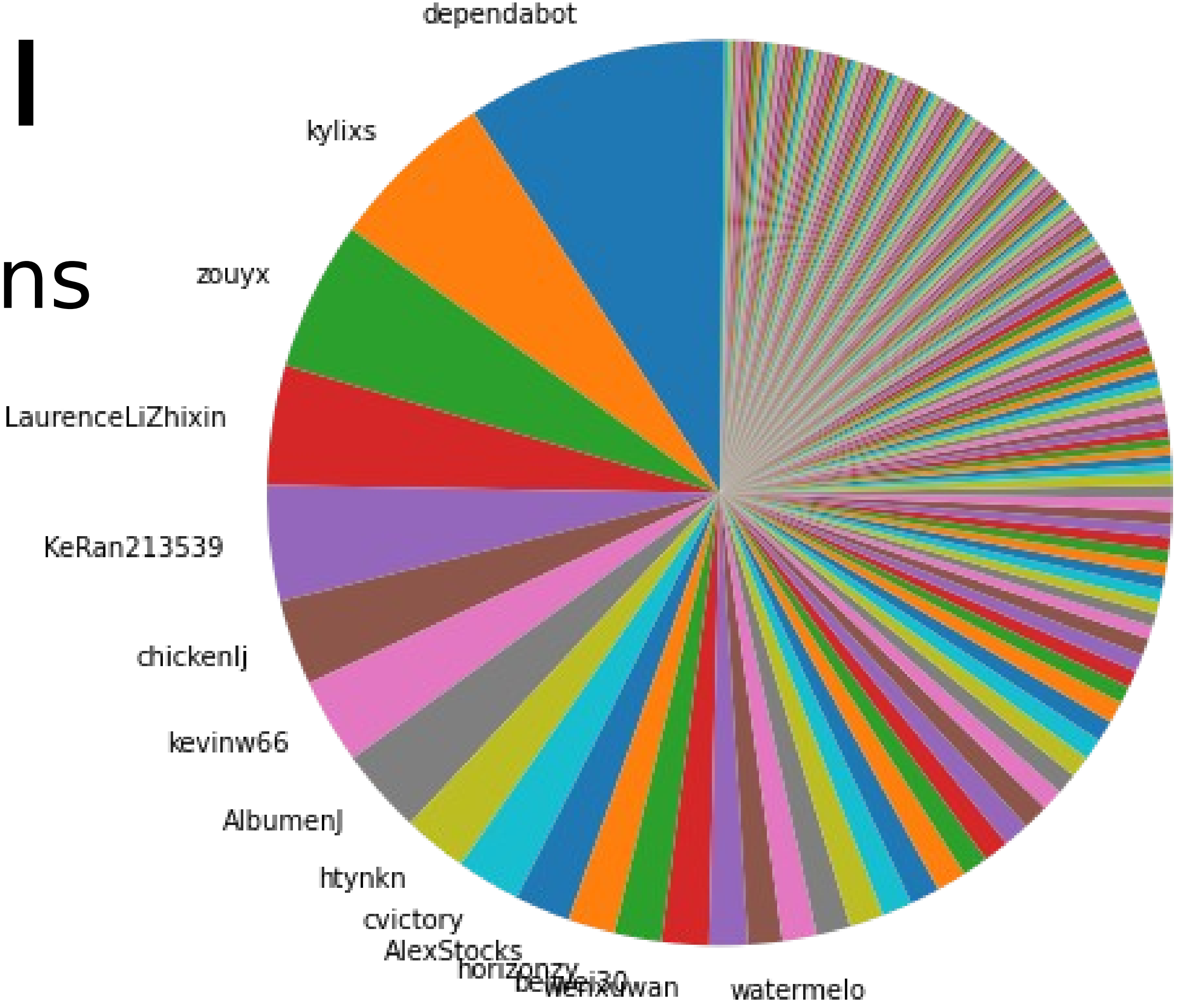


Pony number: 13

First dev ratio: 13.45

The long tail

Small contributions make the calculations unreliable

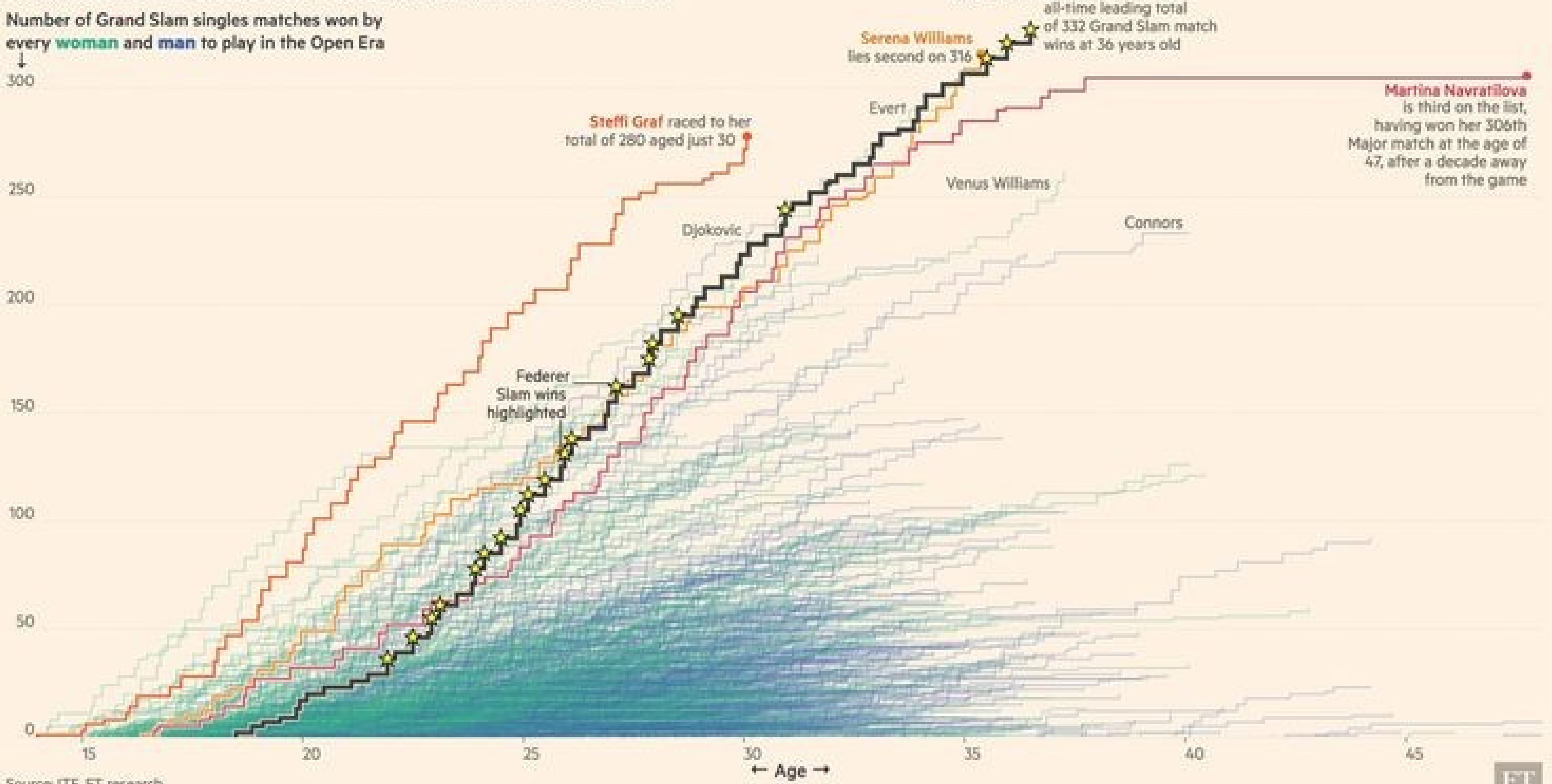


Dubbo

twitter.com/jburnmurdoch/

Federer marches on

In winning the 2018 Australian Open, the 36-year-old Swiss won his 20th Grand Slam title, and extended his all-time record to 332 matches won at tennis' four elite tournaments



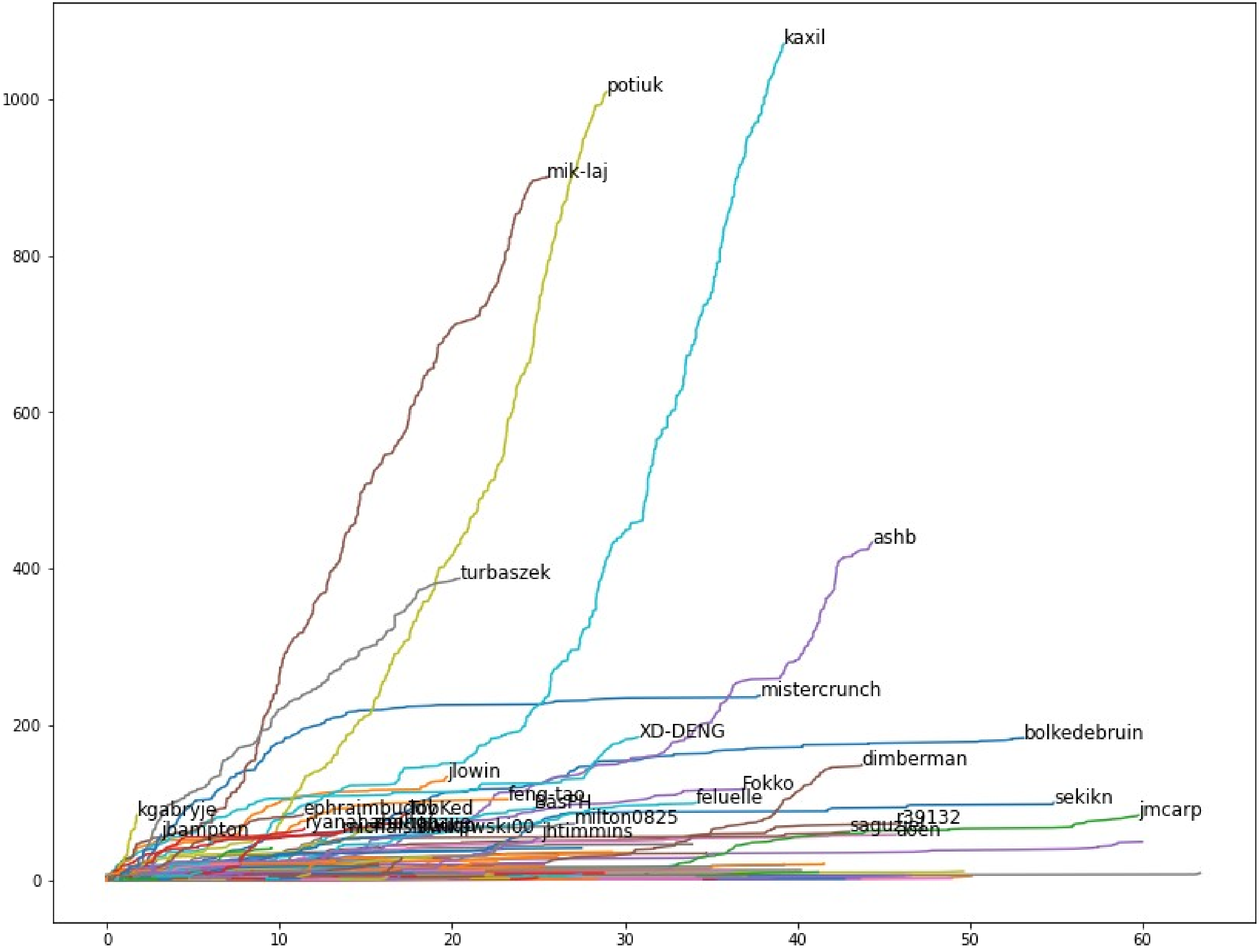
Roger Federer has climbed to the all-time leading total of 332 Grand Slam match wins at 36 years old

Serena Williams lies second on 316

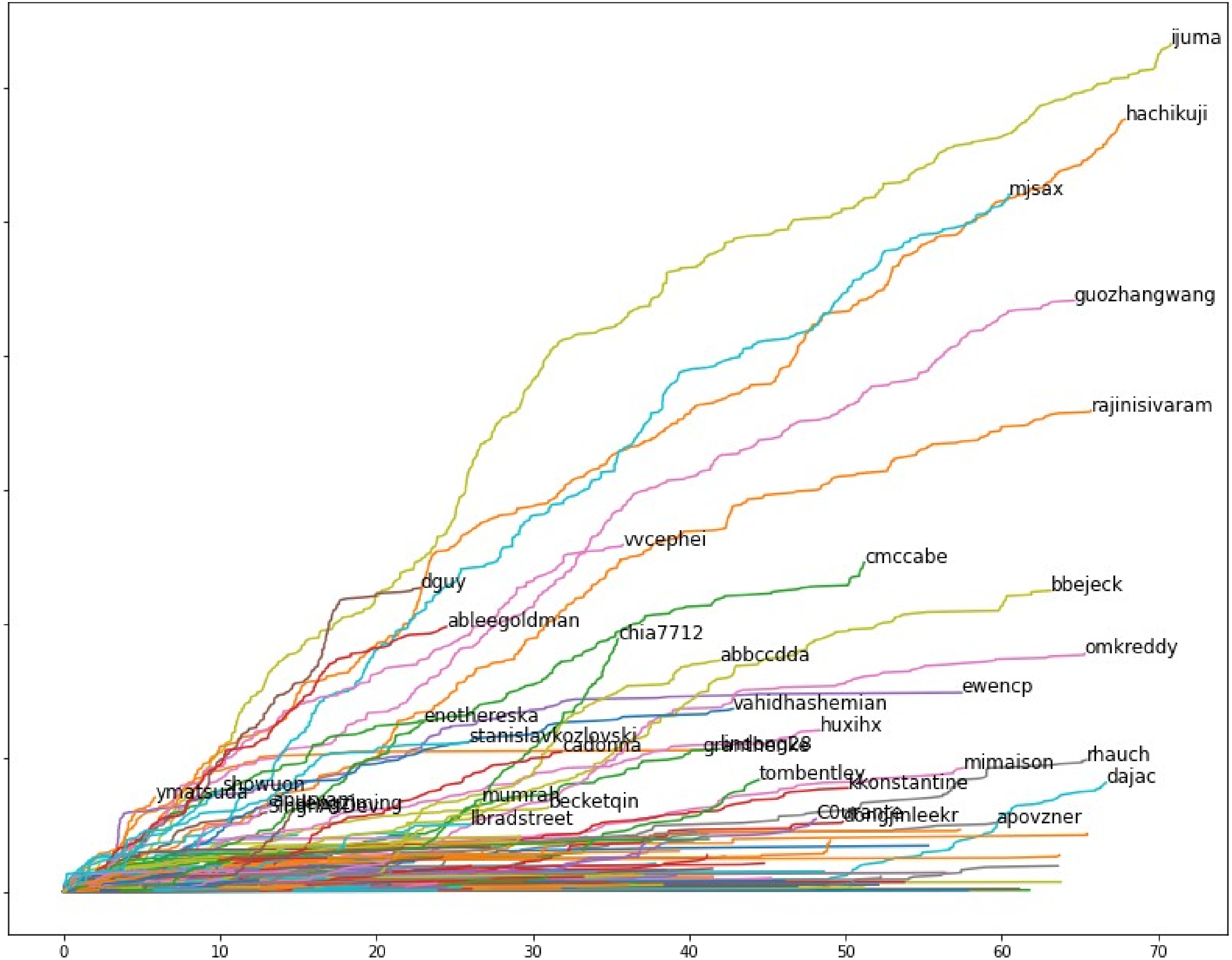
Steffi Graf raced to her total of 280 aged just 30

Martina Navratilova is third on the list, having won her 306th Major match at the age of 47, after a decade away from the game

Source: ITF, FT research
Graphic by John Burn-Murdoch / @jburnmurdoch


kaxil
potiuk
mik-laj
ashb
turbaszek
mistercrunch
XD-DENG
bolkedebruin
dimberman
jlowin
Fokko
feng-tao
feluelle
BasPH
sekikn
jmcarp
kgabryje
ephraimbuddy
Tooked
milton0825
r39132
jbampton
ryanahamilton
saguziel
doen
jhtimmins
mrkm4ntr
michalslowikowski00
1000
800
600
400
200
0
0
10
20
30
40
50
60

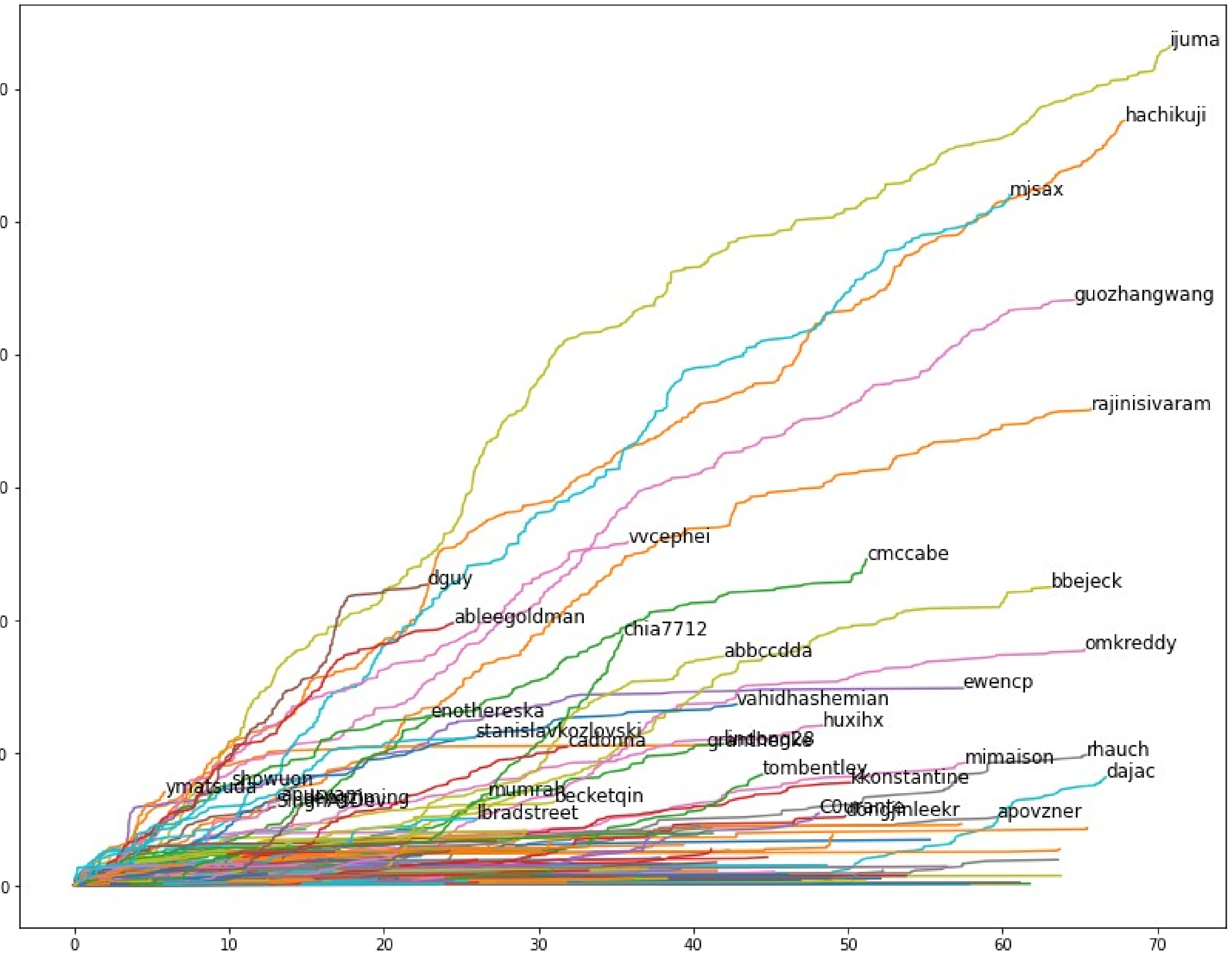
Airflow


ijuma
hachikuji
mjsax
guozhangwang
rajinisivaram
vvcephei
cmccabe
bbejeck
dguy
ableegoldman
chia7712
abbccdda
omkreddy
ewencp
vahidhashemian
enothereska
huxihx
stanislavkozlovski
cadonna
lindong28
granthenke
rhauch
mimaison
tombentley
dajac
kkonstantine
showuon
ymatsuda
apurvam
dengziming
SinghAsDev
mumrah
becketqin
lbradstreet
C0urante
dongjinleekr
apovzner
600
500
400
300
200
100
0
0
10
20
30
40
50
60
70

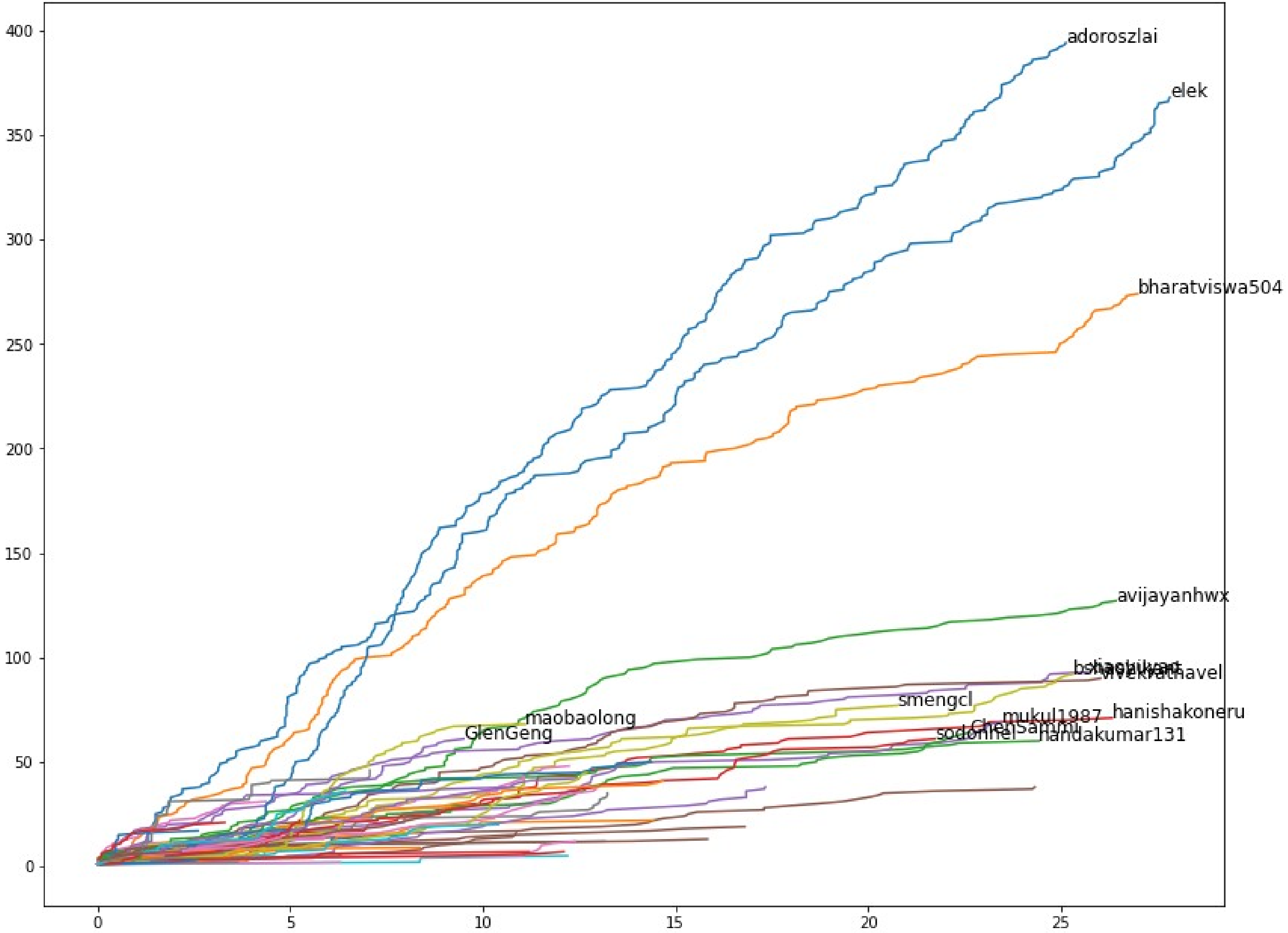
Kafka

The critcal mass

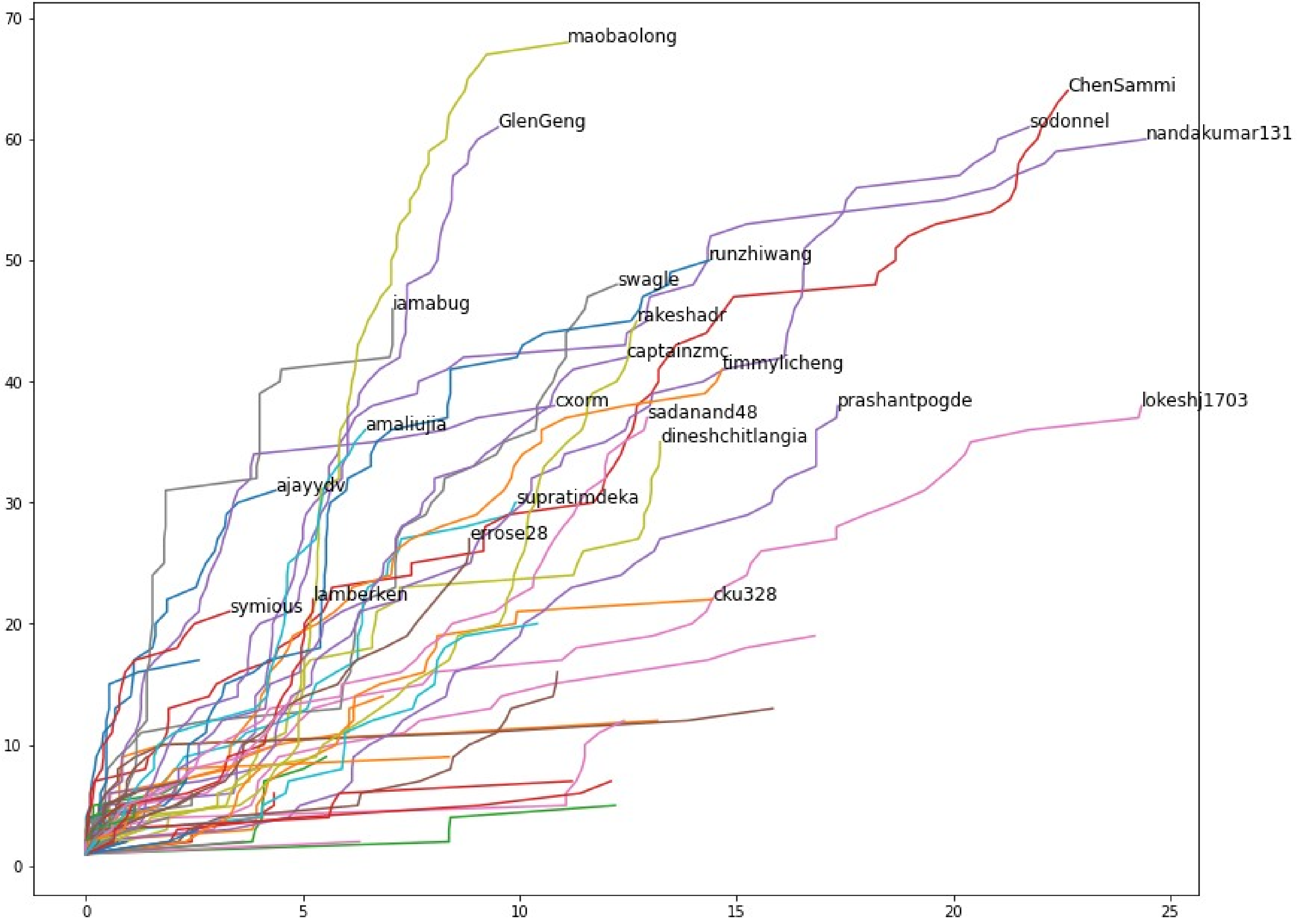
- “every community needs a **group of self-motivated people** with long-term interest, who take responsibility of the project”
- #FOSSback: Maximilian Michels - The Critical Mass: A Guide to Building a Strong Community


0
100
200
300
400
500
600
0
10
20
30
40
50
60
70
ijuma
hachikuji
mjsax
guozhangwang
rajinisivaram
vvcephei
cmccabe
bbejeck
dguy
ableegoldman
chia7712
abbccdda
omkreddy
ewencp
vahidhashemian
enothereska
huxihx
stanislavkozlovski
cadonna
granthenke
lindong28
rhauch
mimaison
tombentley
dajac
kkonstantine
showuon
ymatsuda
SinghAsDev
mumrah
becketqin
lbradstreet
C0urante
dongjinleekr
apovzner

Kafka

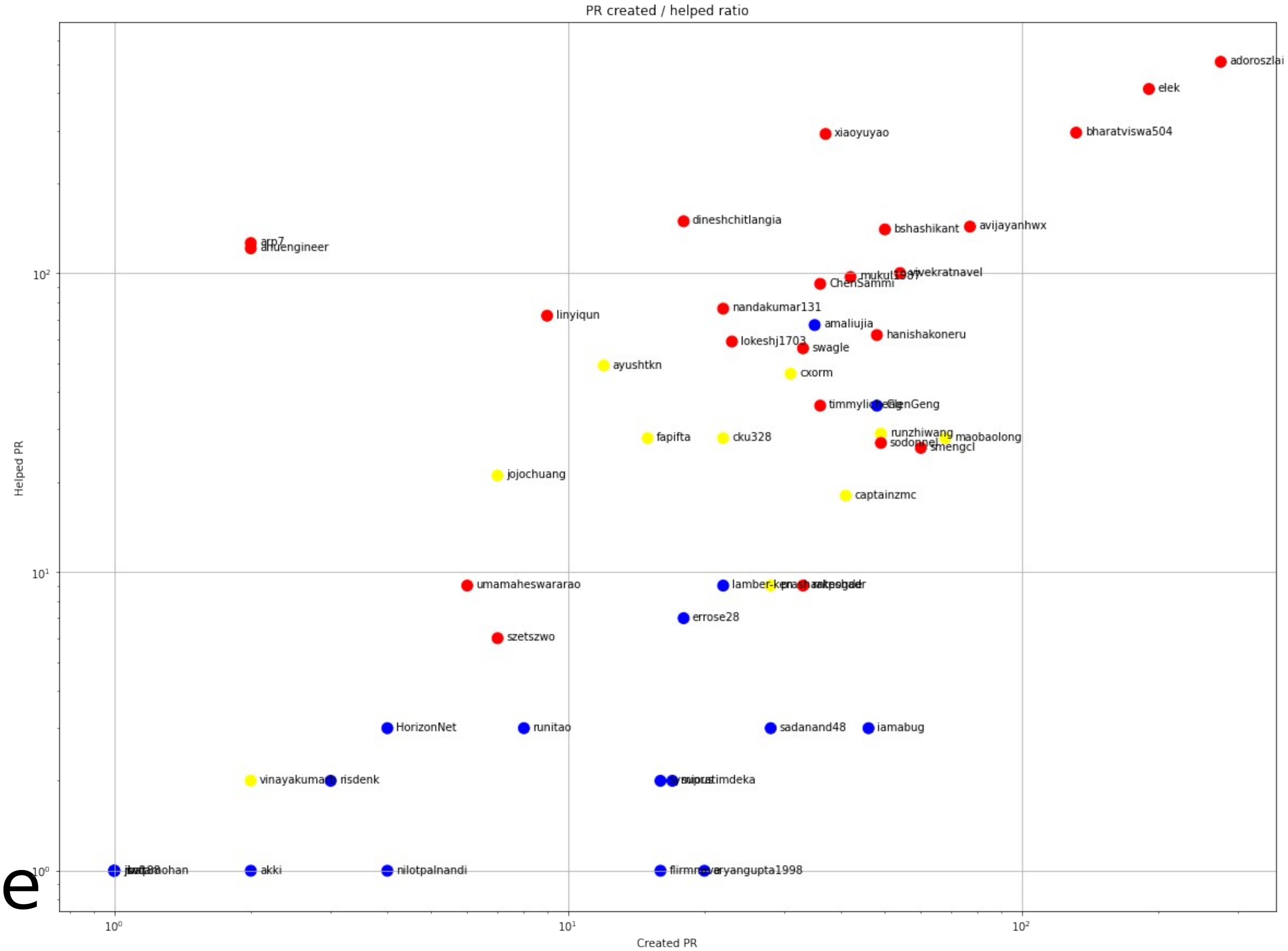

adoroszlai
elek
bharatviswa504
avijayanhwx
bshviyoshikwv
vivekratnavel
smengcl
maobaolong
GlenGeng
mukul1987
hanishakoneru
ChenSammi
sodonnel
nandakumar131
400
350
300
250
200
150
100
50
0
0
5
10
15
20
25

Ozone

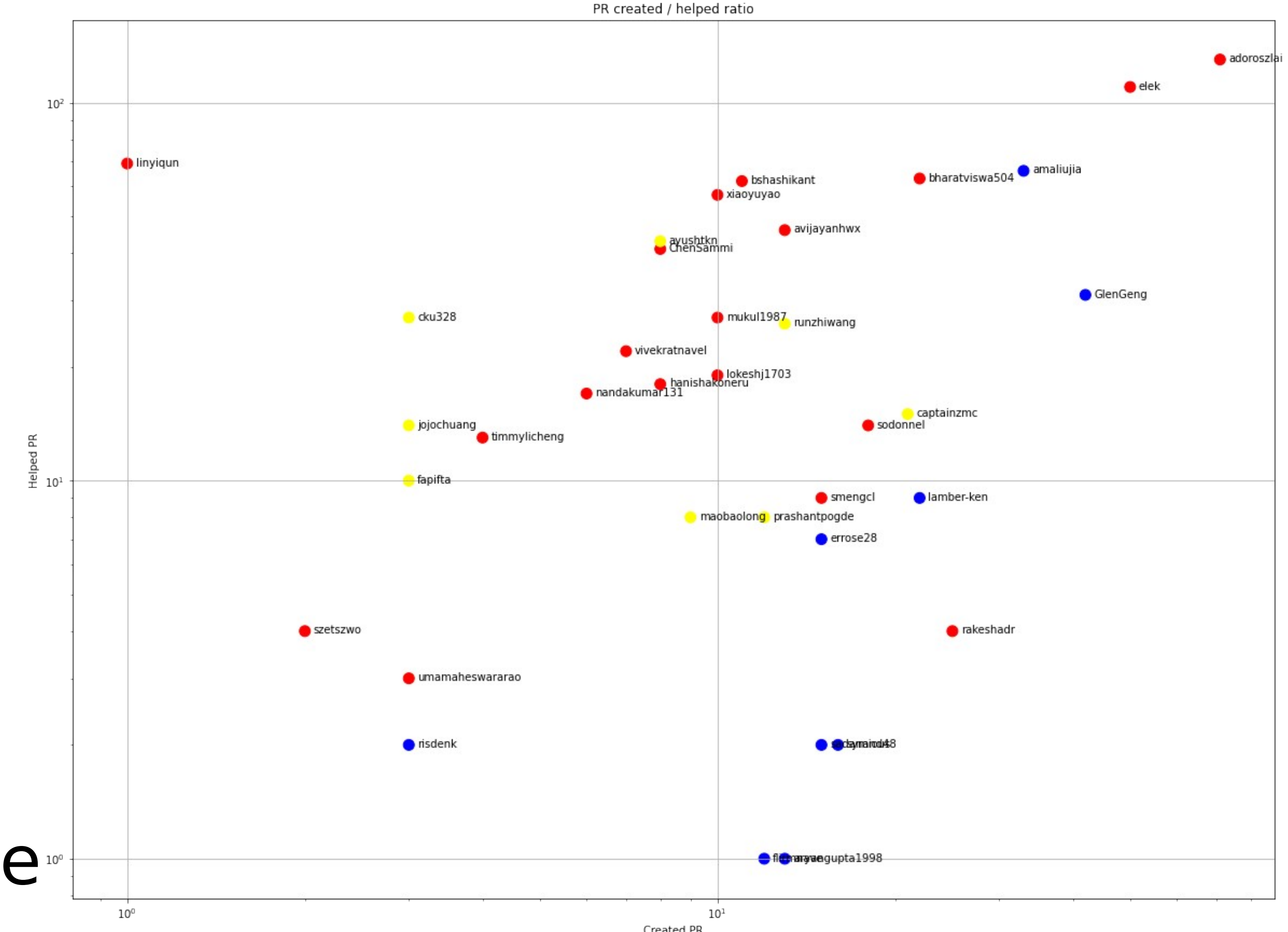

maobaolong
ChenSammi
GlenGeng
sodonnel
nandakumar131
runzhiwang
swagle
iamabug
rakeshadr
captainzmc
timmylicheng
cxorm
prashantpogde
lokeshj1703
sadanand48
amaliujia
dineshchitlangia
ajayydv
supratimdeka
errose28
cku328
lamberken
symious
70
60
50
40
30
20
10
0
0
5
10
15
20
25

Ozone

Actions behind numbers?

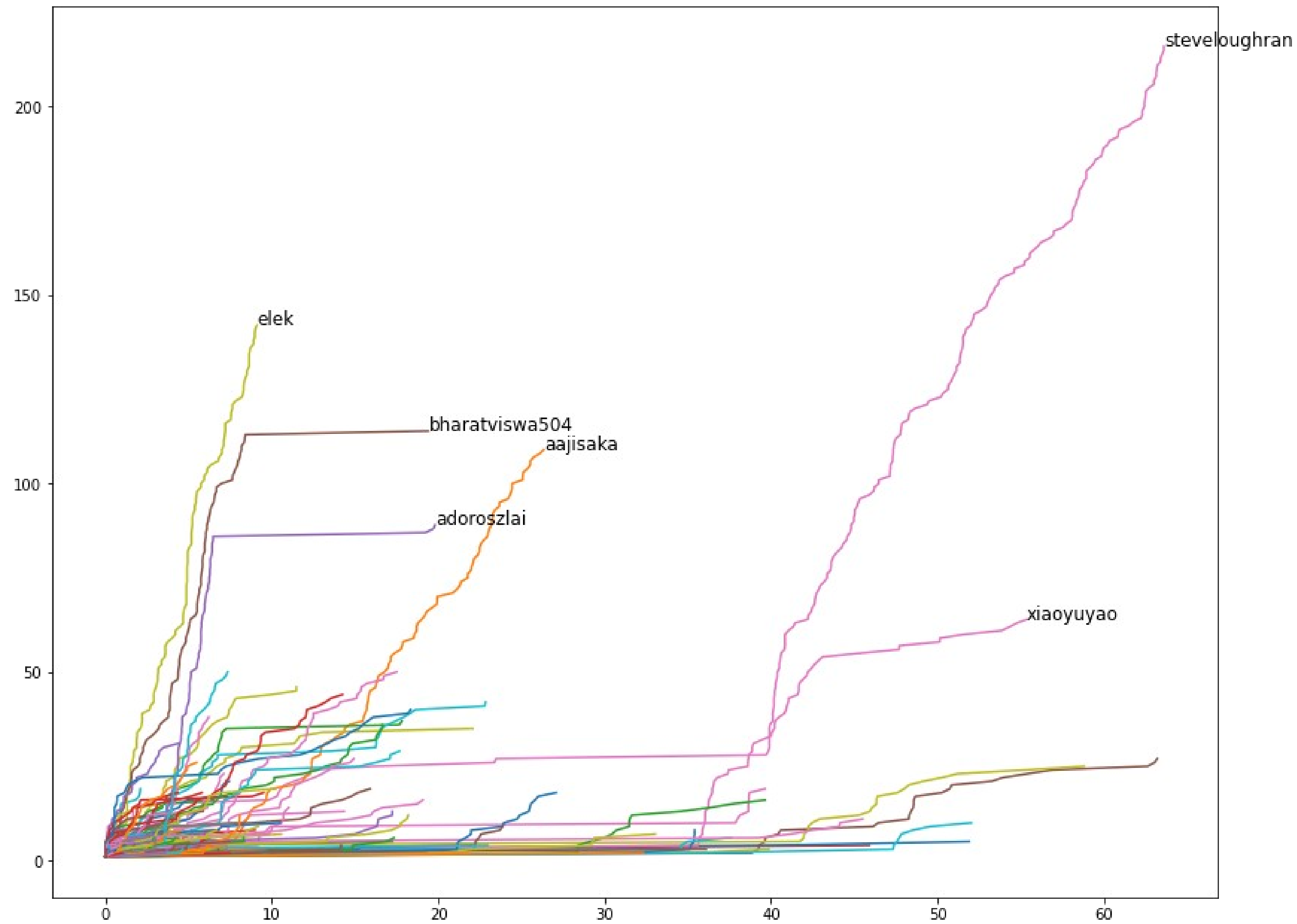

PR created / helped ratio
Helped PR
Created PR
10^0
10^1
10^2
adoroszlai
elek
bharatviswa504
xiaoyuyao
dineshchitlangia
bshashikant
avijayanhwx
arp7
ahuengineer
mukul1987
vivekratnavel
ChenSammi
nandakumar131
linyiqun
amaliujia
hanishakoneru
lokeshj1703
swagle
ayushtkn
cxorm
timmyli
GlenGeng
fapifta
cku328
runzhiwang
maobaolong
sodonnel
smengcl
jojochuang
captainzmc
umamaheswararao
lamber-ke
errose28
szetszwo
HorizonNet
runitao
sadanand48
iamabug
vinayakuma
risdenk
timdeka
akki
nilotpalnandi
flirmn
aryangupta1998

Ozone


PR created / helped ratio
adoroszlai
elek
linyiqun
amaliujia
bharatviswa504
bshashikant
xiaoyuyao
avijayanhwx
ayushtkn
ChenSammi
GlenGeng
cku328
mukul1987
runzhiwang
vivekratnavel
lokeshj1703
hanishakoneru
nandakumar131
captainzmc
sodonnel
jojochuang
timmylicheng
fapifta
smengcl
lamber-ken
maobaolong
prashantpogde
errose28
szetszwo
rakeshadr
umamaheswararao
risdenk
[illegible]
[illegible]gupta1998
10^2
10^1
10^0
Helped PR
10^0
10^1
Created PR

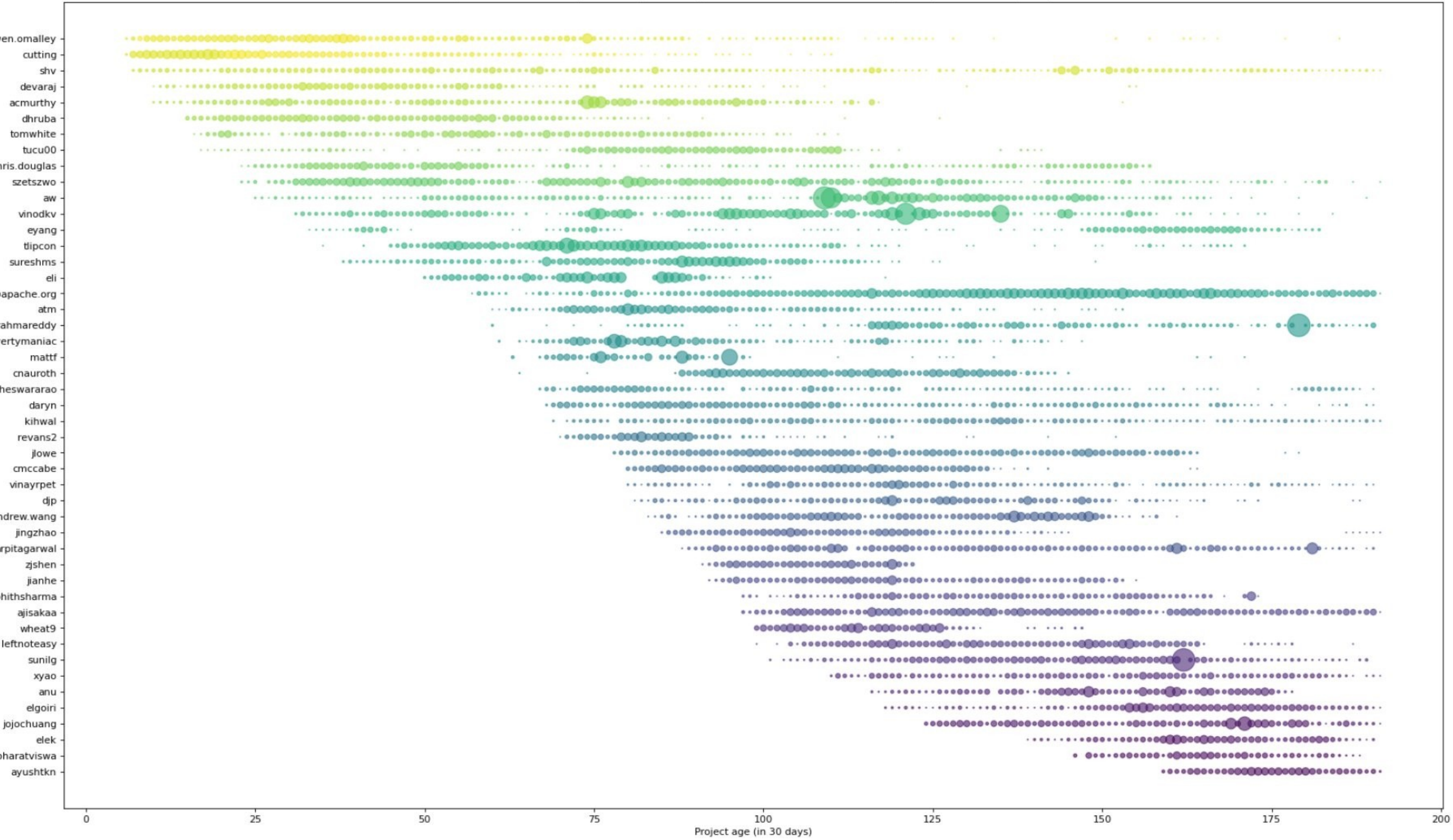
Ozone

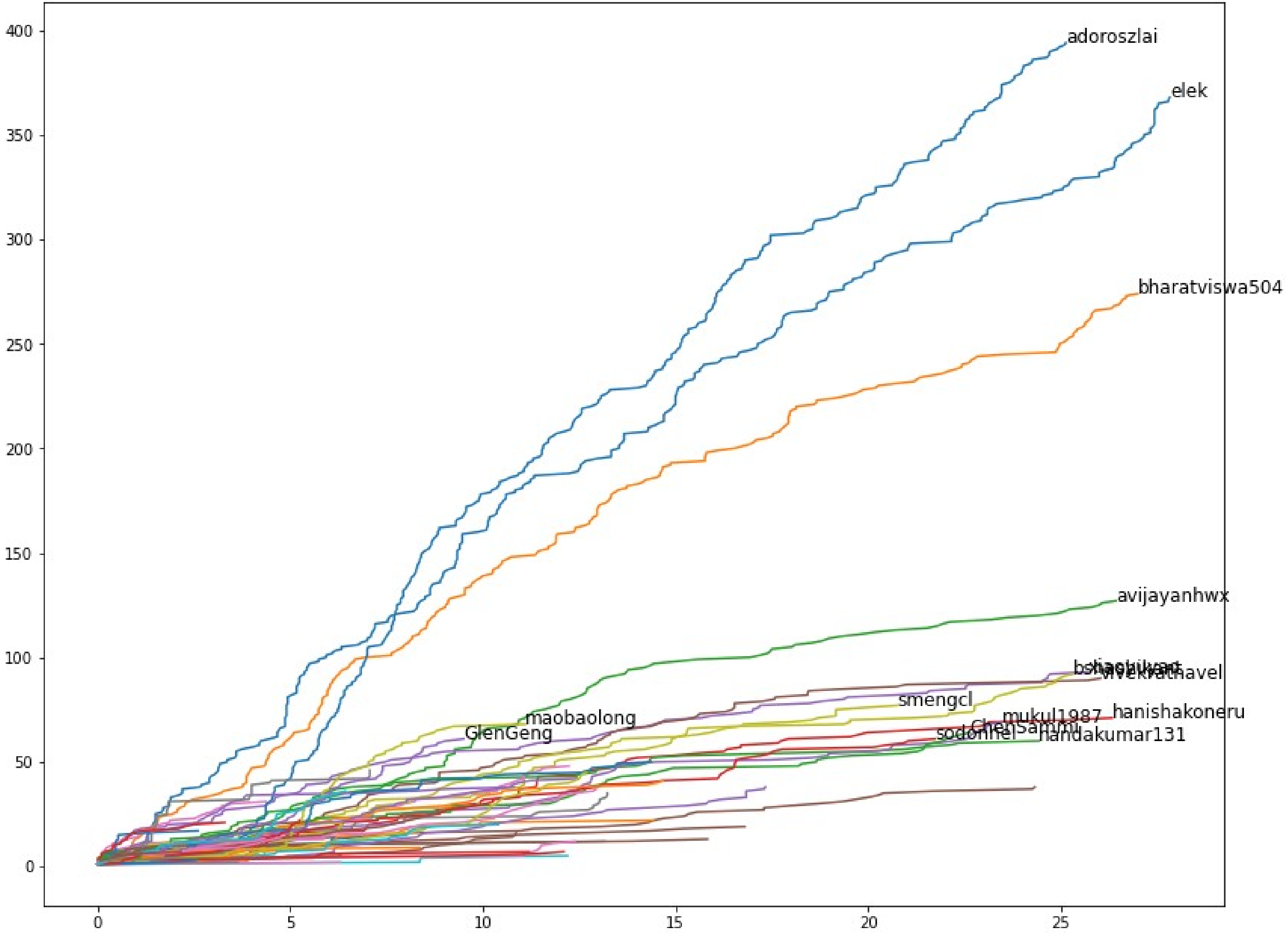
Time, again


stevelouughran
elek
bharatviswa504
aajisaka
adoroszlai
xiaoyuyao
200
150
100
50
0
0
10
20
30
40
50
60

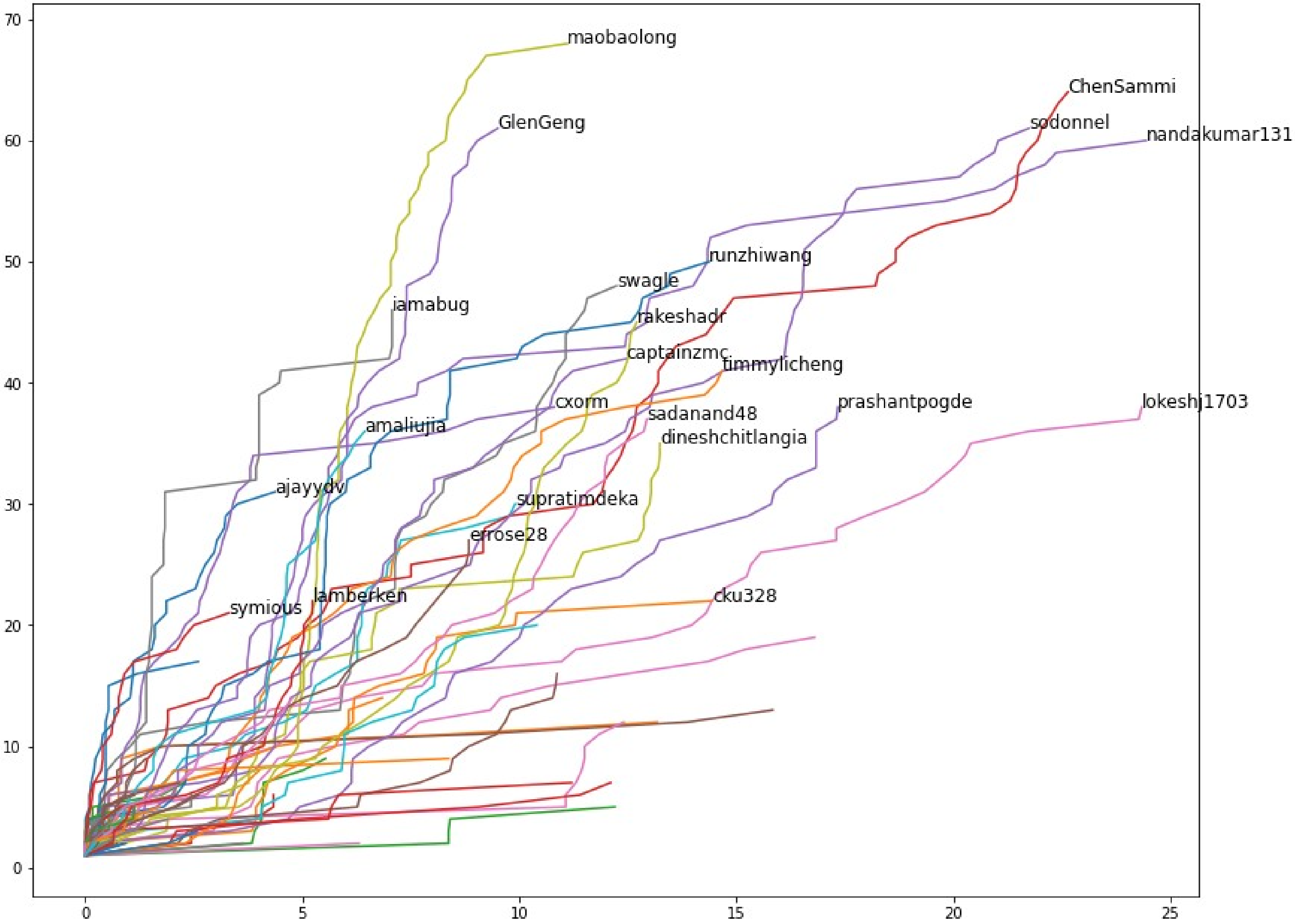
Hadoop

Apache Hadoop contributors contributed to the most Jira issues
(dots are number of Jira issues per 30 days where the contributor had any work, comment or patch)



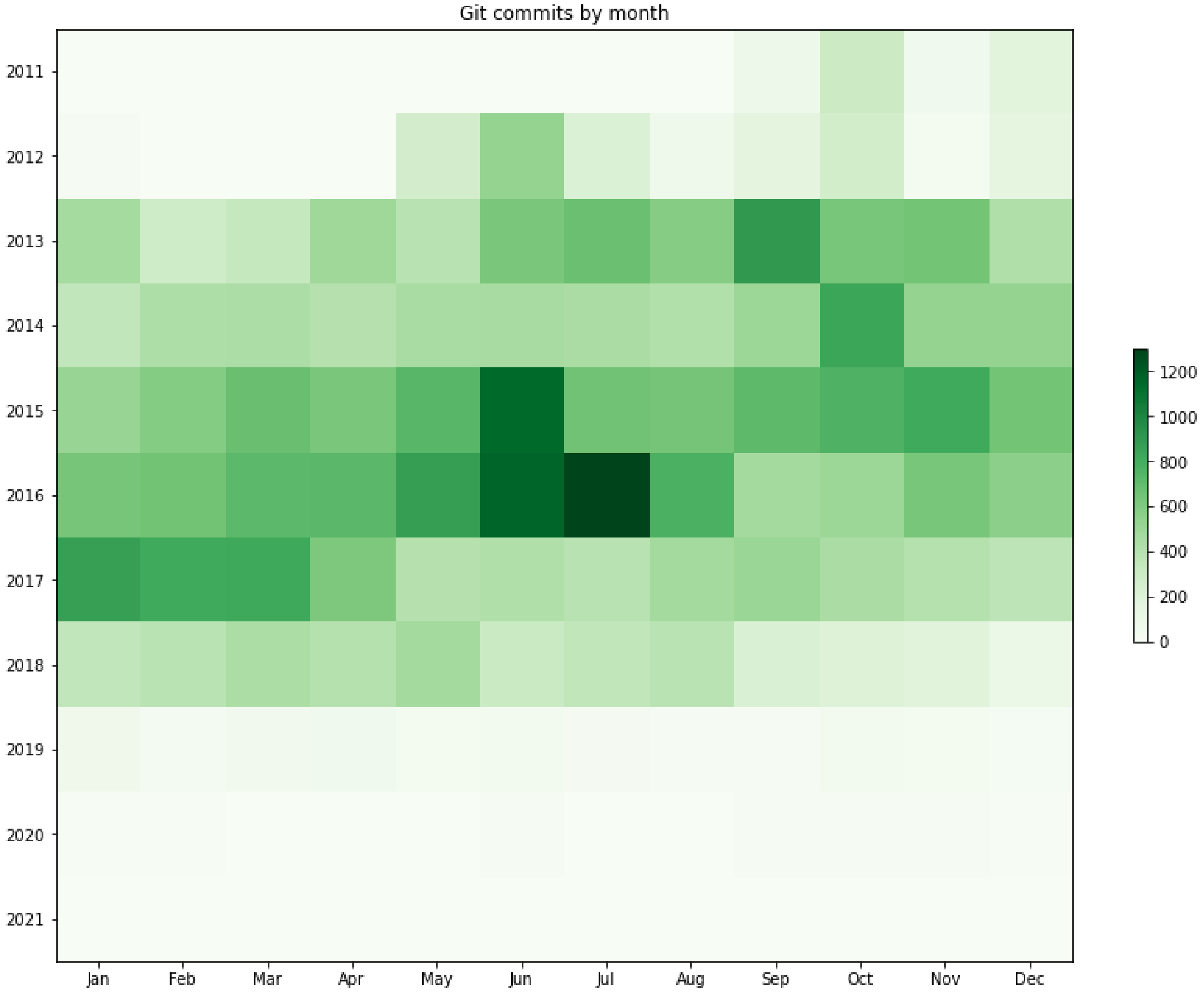

adoroszlai
elek
bharatviswa504
avijayanhwx
bshiaobikuaq
vivekrathavel
smengcl
maobaolong
GlenGeng
mukul1987
hanishakoneru
ChenSammi
sodonnel
nandakumar131
400
350
300
250
200
150
100
50
0
0
5
10
15
20
25

Ozone



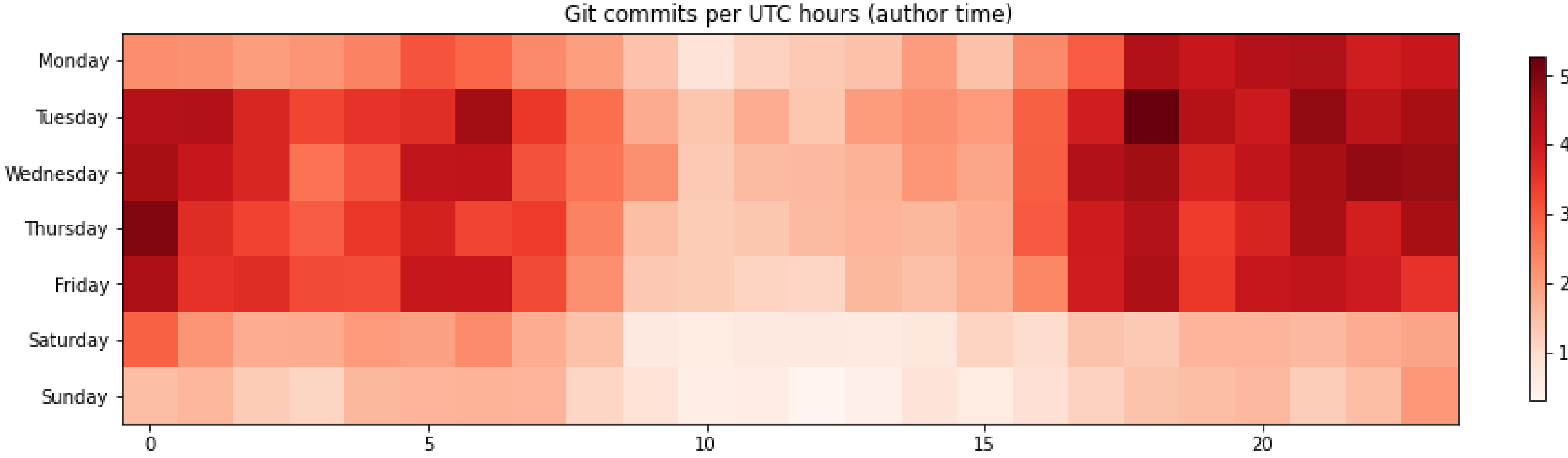
maobaolong
ChenSammi
GlenGeng
sodonnel
nandakumar131
runzhiwang
swagle
iamabug
rakeshadr
captainzmc
timmylicheng
cxorm
prashantpogde
lokeshj1703
sadanand48
amaliujia
dineshchitlangia
ajayydv
supratimdeka
errose28
symious
lamberken
cku328
0
10
20
30
40
50
60
70
0
5
10
15
20
25

Ozone

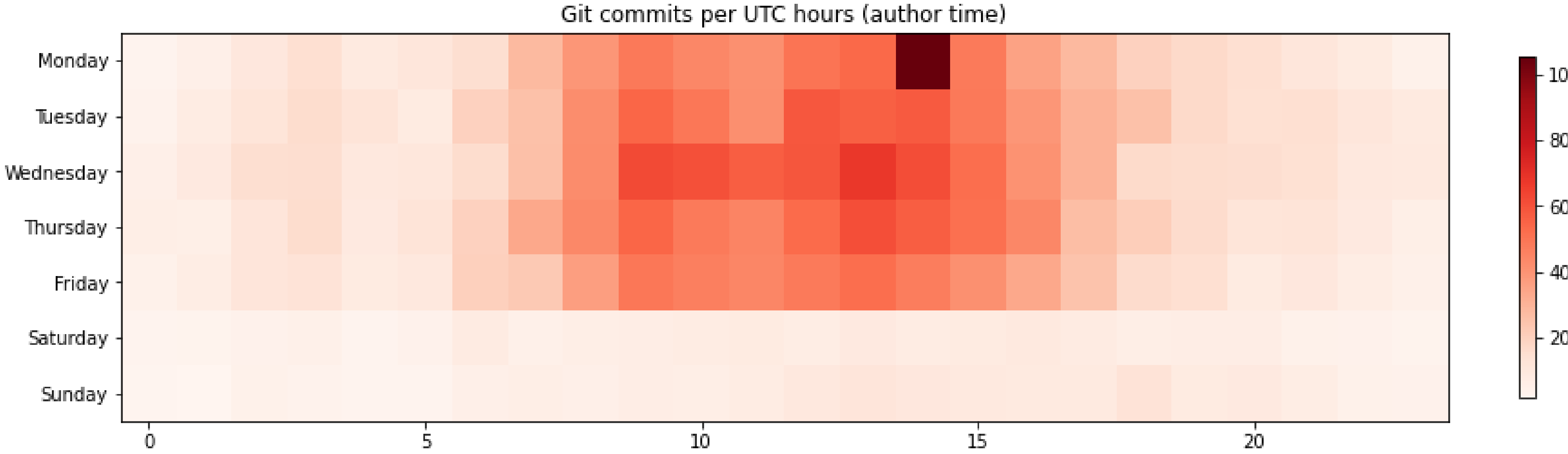


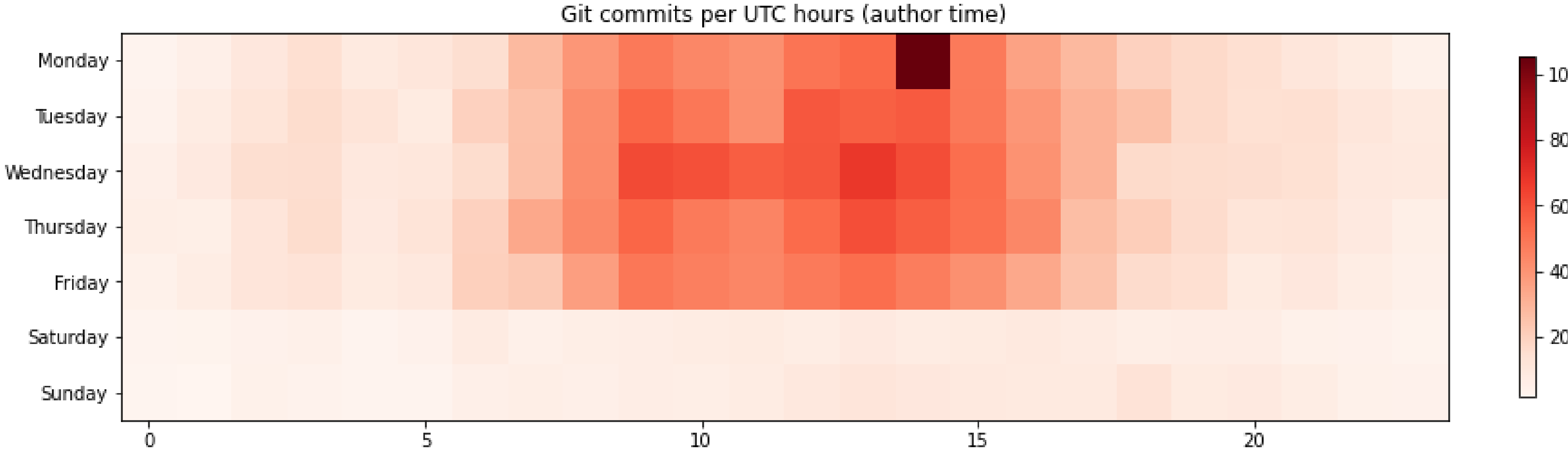
Ambari

“every community needs a group of self-motivated people ...”

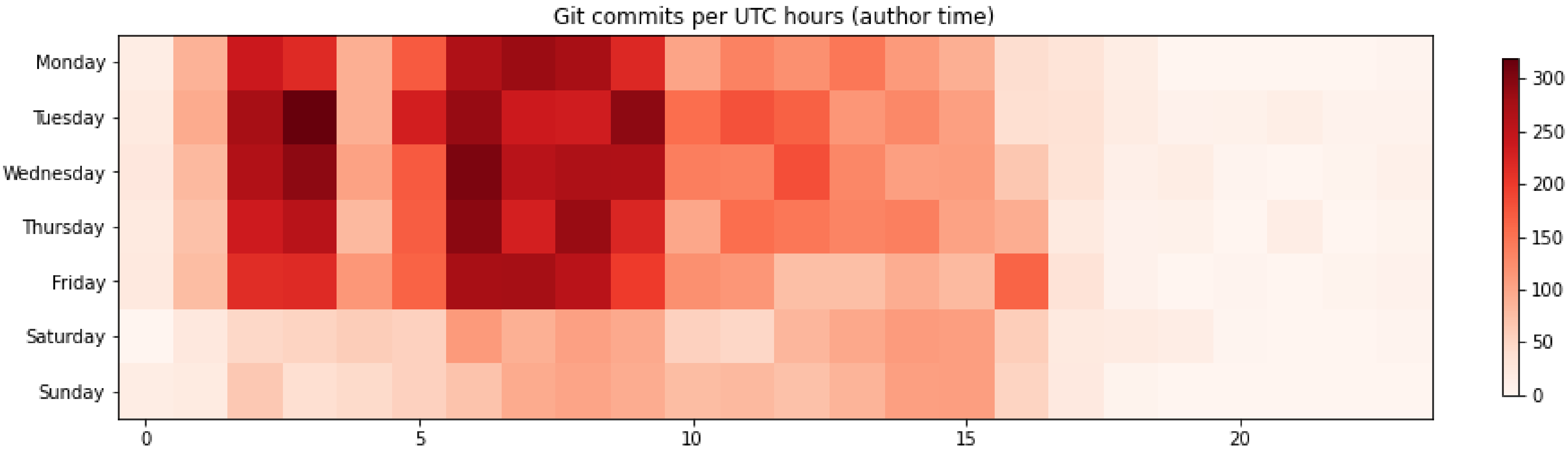

Git commits per UTC hours (author time)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
0
5
10
15
20
5
4
3
2
1

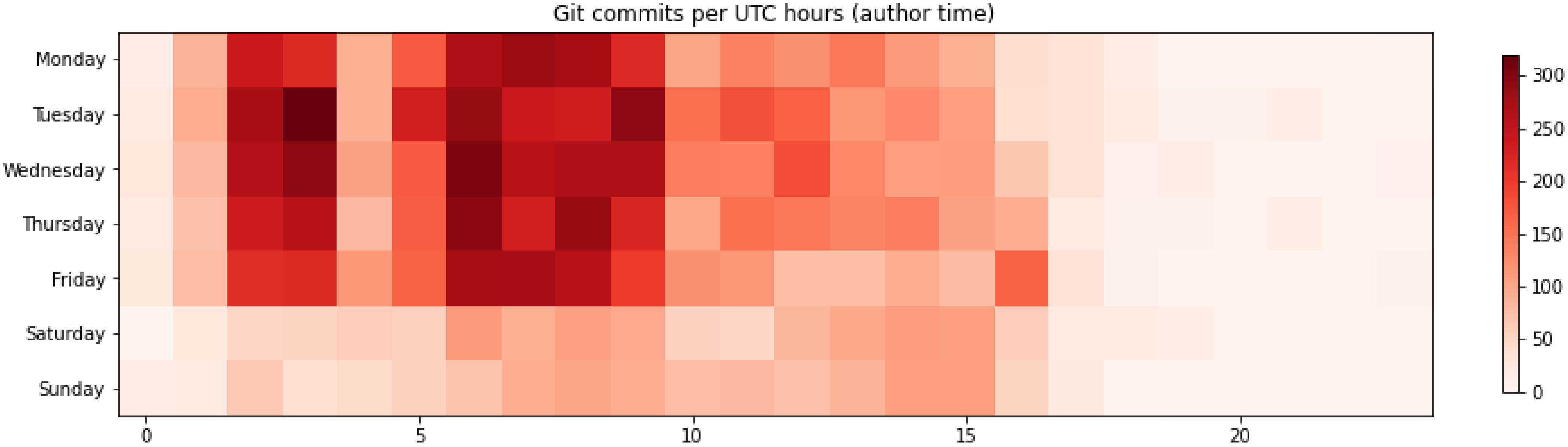
Spark


Git commits per UTC hours (author time)

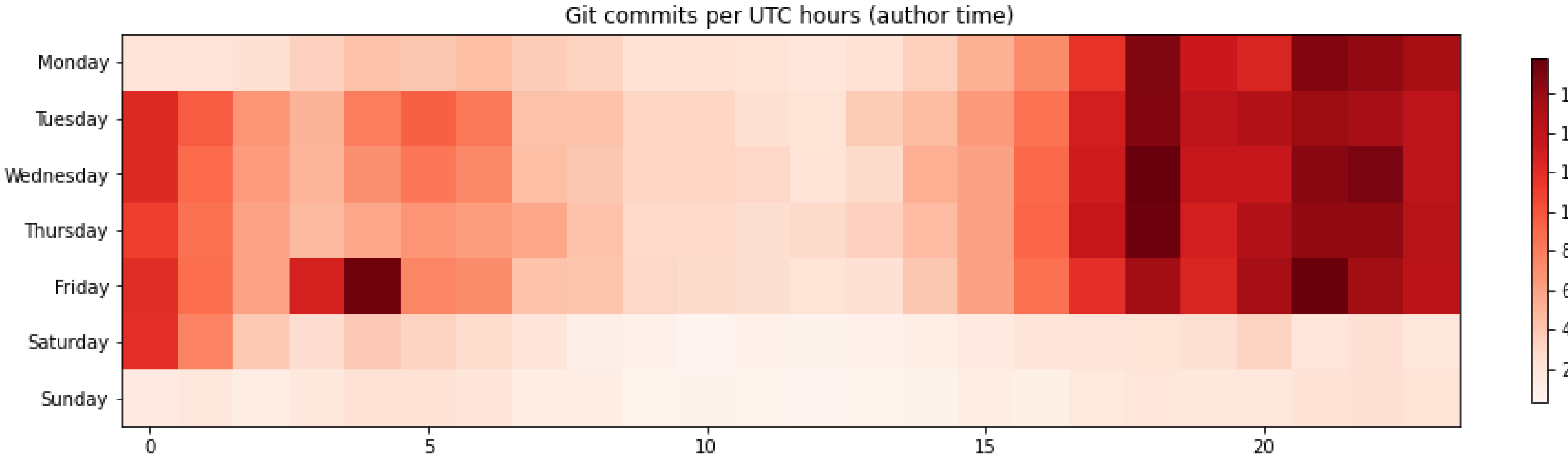

Git commits per UTC hours (author time)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
0
5
10
15
20
10
80
60
40
20

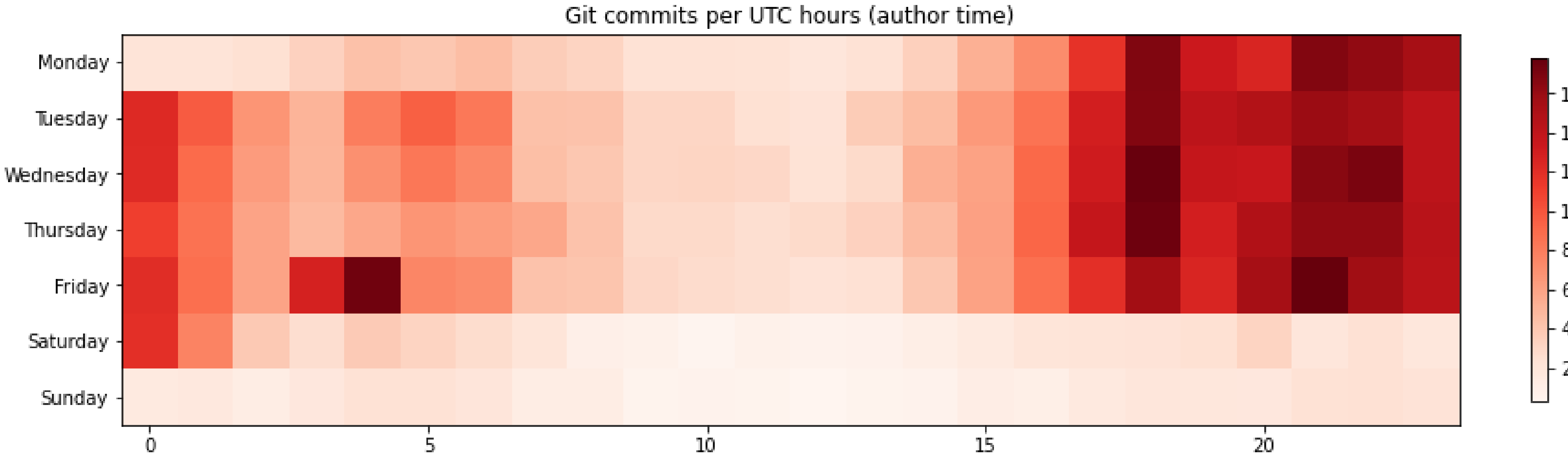
Flink


Git commits per UTC hours (author time)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
0
5
10
15
20
300
250
200
150
100
50
0

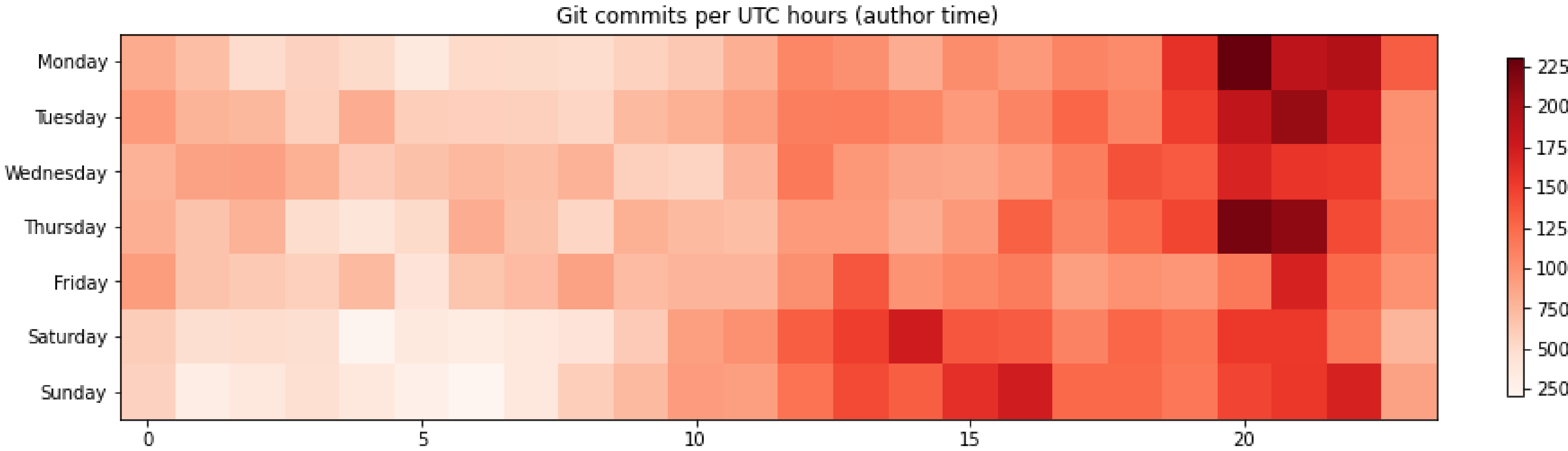

Git commits per UTC hours (author time)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
0
5
10
15
20
300
250
200
150
100
50
0

Dubbo

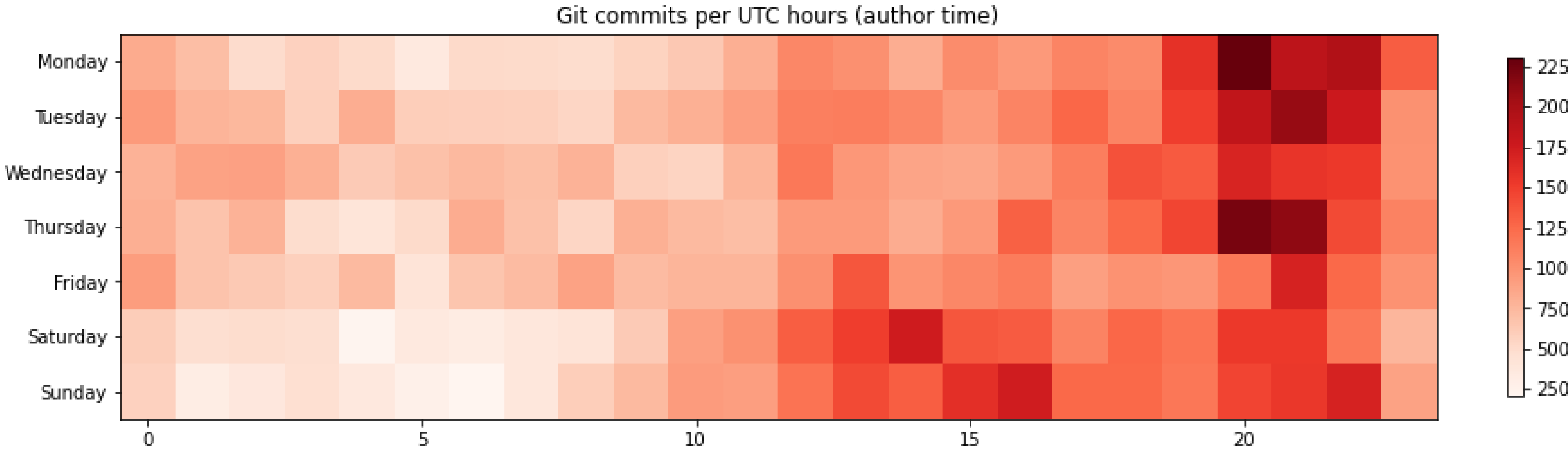

Git commits per UTC hours (author time)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
0
5
10
15
20
1
1
1
1
8
6
4
2


Git commits per UTC hours (author time)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
0
5
10
15
20
1
1
1
1
8
6
4
2

Hadoop



Git commits per UTC hours (author time)


Git commits per UTC hours (author time)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
0
5
10
15
20
225
200
175
150
125
100
750
500
250

Maven

Summary?

- Any statistic just a small window to the real world
- One number couldn’t tell the story (bus factor)
- Trend!

Summary?

- Any statistic just a small window to the real world
- One number couldn’t tell the story (bus factor)
- Trend!

…one day I might get hit by a bus or get cancer

But right now all I am is a fabulous dancer…

so dance with me baby...

(The Burning Hell)

Márton Elek

- elek@apache.org
- github.com/elek
  - https://github.com/elek/bus-factor
  - https://github.com/elek/asf-project-stat
- twitter.com/@anzix
- Code and Console (Youtube)
- Ozone Explained (Youtube)