**Typology of repos**

To get this to work within bigquery’s limitations I’ve had to split the data into chunks. I’m doing this by repo creation date – the first chunk I’m working with only considers repos created after 1st June 2013 until whenever the timeline was copied into our github\_explore folder (about 1.4 million repos).

Types 1-10 are ‘asocial’ repos – I’ve defined this as having less than 2 Watchers or Forks and no PullRequests

Types 1-5 were created from scratch, 6-10 apply the same criteria to repos which are forks

I then have 5 types of repo related to pull requests - concerning whether the repo was issuing or receiving pull requests (or both).

The final 2 types are social repos – they have watchers and/or forks – and these take precedence over pull requests (if the repo receives pull requests but also has watchers/forks its type 16/17)

**Simple typology:**

1 – Pushes only, single user

2 – Pushes only, multiple users

3- Fork with Single Pusher

4 – Fork with multiple pushers

5- Pull Request Issuer

6- Pull Request Receiver

7 – Has Social Events (>1 Watchers or Forks)

8 – More than 50 Watchers or Forks

**Comprehensive Typology**

**Pushes only, not a fork**

Type 1 – single push

Type 2 – Short term personal use - Pushes only, active for less than 1 month, single pusher

Type 3 – Short term collaboration - Pushes only, active for less than 1 month, more than one pusher

Type 4 – Medium term personal use – Pushes only, active for more than 1 month, single pusher

Type 5 - Medium term collaboration - Pushes only, active for more than 1 month, more than one pusher

**Forks without social events**

Type 6 – Single Push

Type 7 – Short term personal use - Pushes only, active for less than 1 month, single pusher

Type 8 – Short term collaboration - Pushes only, active for less than 1 month, more than one pusher

Type 9 – Medium term personal use – Pushes only, active for more than 1 month, single pusher

Type 10 - Medium term collaboration - Pushes only, active for more than 1 month, more than one pusher

**Pull Request repos**

Type 11 – Issues Single Pull Request

Type 12 – Issues Multiple Pull Requests

Type 13 – Receives and Issues pull requests but no other social events

Type 14 – Receives Pull Requests but no other social events

Type 15 - Pushes continue 1 month after pull requests cease

**Social Repos**

Type 16 – slightly social - 2-50 Watchers or Forks

Type 17 – well known – 51+ Watchers or Forks

When these types are applied to the **first chunk** of data it results in the following frequencies:

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Count | Not Orgs | Organisations |
| 1 | 262668 | 18.63% | 14.48% |
| 2 | 555799 | 39.69% | 27.8% |
| 3 | 45048 | 2.7% | 7.6% |
| 4 | 114231 | 8% | 8% |
| 5 | 18685 | 1% | 5% |
| 6 | 67217 | 4.5% | 6% |
| 7 | 74396 | 5% | 6.5% |
| 8 | 2637 | 0.1% | 1% |
| 9 | 13423 | 0.9% | 1.3% |
| 10 | 1698 | 0.05% | 0.95% |
| 11 | 143957 | 10.5% | 4.4% |
| 12 | 58262 | 4.3% | 1.5% |
| 13 | 5882 | 0.4% | 0.3% |
| 14 | 6150 | 0.3% | 1.4% |
| 15 | 1135 | 0.08% | 0.06% |
| 16 | 64903 | 3.7% | 13.1% |
| 17 | 1825 | 0.1% | 0.4% |

Re-produce tables to get org values

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Count | Not Orgs | Organisations |
| Orphan | 1779831 |  |  |
| ForkOrphan | 2630081 |  |  |

Could add in forks that were never pushed to