https://github.com/datactive/bigbang/

BigBang @ IETF 110 Hackathon

Sebastian Benthall

What is BigBang

- A scientific toolkit for studying collaborative communities
- Data sources: Email, Git repositories, IETF DataTracker, ListServ, …
- Data science tools: using SciPy stack
  - Entity resolution for names and organizations
  - Social network analysis
  - Natural language processing on message content
  - Time series analysis
  - Information extraction...

What about arkko.com/tools/rfcstats/ ?

- We love rfcstats and are inspired by it.
- BigBang uses a wider range of data sets beyond the IETF Datatracker, such as mailing lists.
- It supports different kinds of research questions.
- BigBang developers/users tend to be either:
  - Social scientists studying standardization and/or collaboration
  - Computer scientists developing new data science methods

UNIVERSITY OF AMSTERDAM

Berkeley
UNIVERSITY OF CALIFORNIA

DATACTIVE

Outcomes from IETF 110 Sprint: Software community

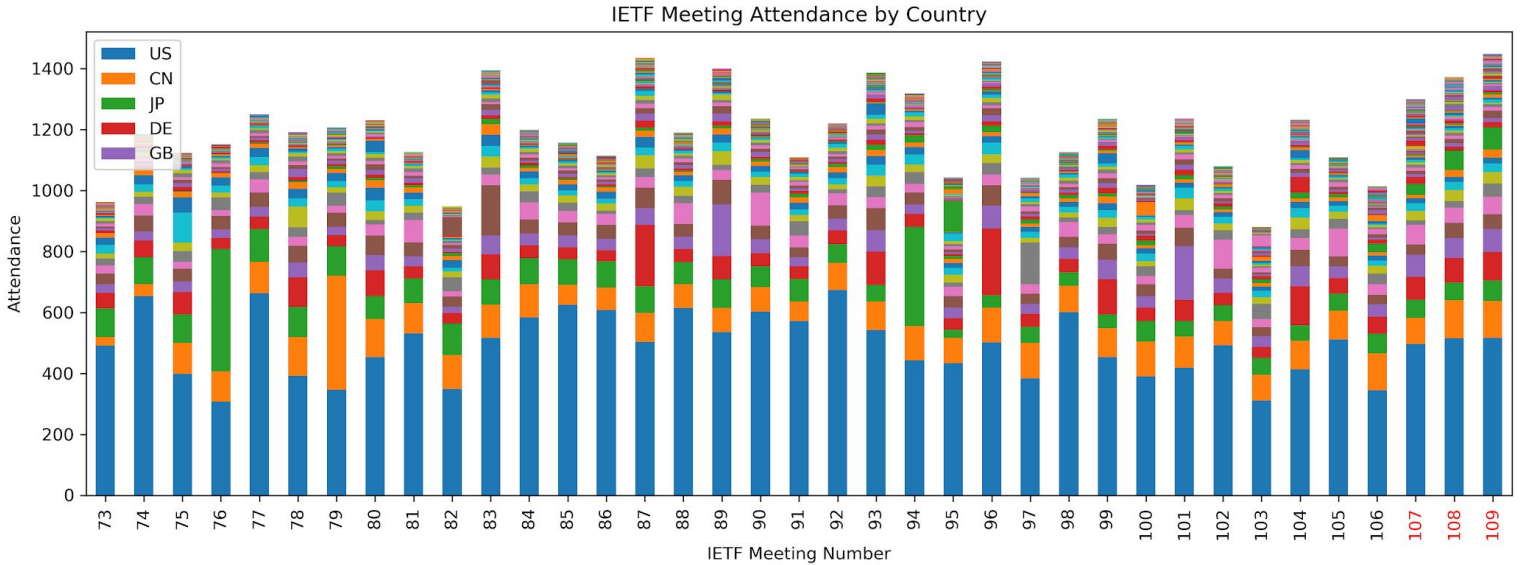
- *Growth.* New participants in the project!
- *Maintenance*. Updated installation instructions to keep up with dependencies.
- *Onboarding*. Produced instructional videos for installation and basic usage.
- *Debugging*. Debugged ingest issues around malformed data.
- *New data sources*. Work towards scraping Listserv; used by other standards organizations such as 3GPP.

Outcomes from IETF 110 Sprint: Science!

- *Attendance analysis*. Impact of remote meetings on IETF 110 attendance. (Nick Doty)

- *Organizational involvement*. Building tools to better understand the involvement of organizations in IETF and other standards groups.

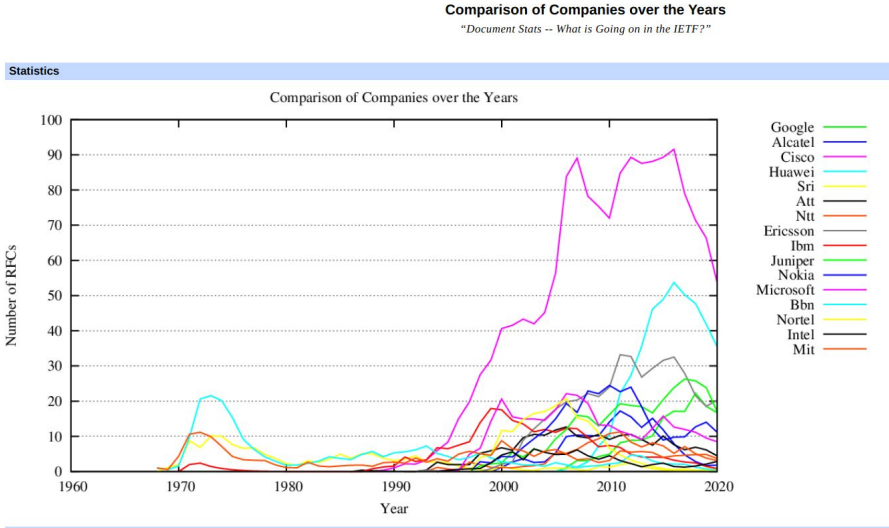
Remote meetings and attendance



The virtual meetings have modestly higher attendance than recent meetings. The proportions by country are not obviously different in the virtual meetings, but there may be less variation of the proportion of attendance based on where the meeting is physically located. (That is, so far we don't see the big swings in US, Chinese, Japanese or German attendance, as we did when the meeting was physically located in the US, China, Japan or Europe.) [Nick Doty]

Organizational involvement

- Research interest in which organizations are influential in which working groups.
- Datatracker/authorstats is a great resource for this
- But this does not generalize to other standards groups
- We are exploring analysis using email domains



Working with email domains

- myname@myorg.tld -- a way to identify an org's role on a mailing list.

- Challenges:
  - Individuals with personal email domains.
  - Generic email hosting domains -- e.g. gmail.com, gmx.de, etc.

- Threshold on entropy of distribution of email addresses per domain filters out personal domains.

$$H(D) = -\sum_{e \in D} \frac{n_e}{n_D} \log \frac{n_e}{n_D}$$

- Still working on a solution for generic email hosts.

Future plans

- New release with improved documentation
- Containerized environment for IETF data exploration using interactive notebooks
- Refactoring core code for better encapsulation
- Complete organizational involvement analysis for IETF and compare with other standards groups such as 3GPP, W3C, ICANN, ...
- Integration with information extraction toolkits for knowledge graph construction

https://github.com/datactive/bigbang/

To learn how to contribute and join the mailing list, check the README!