Towards a Data Science of Institutional Power: Progress with BigBang

Human Rights Protocol Consideration @ IETF 110

Sebastian Benthall

This talk

- What is BigBang?
- BigBang @ IETF 110
- Vision and Strategy: Data Science of Institutional Power

What is BigBang

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

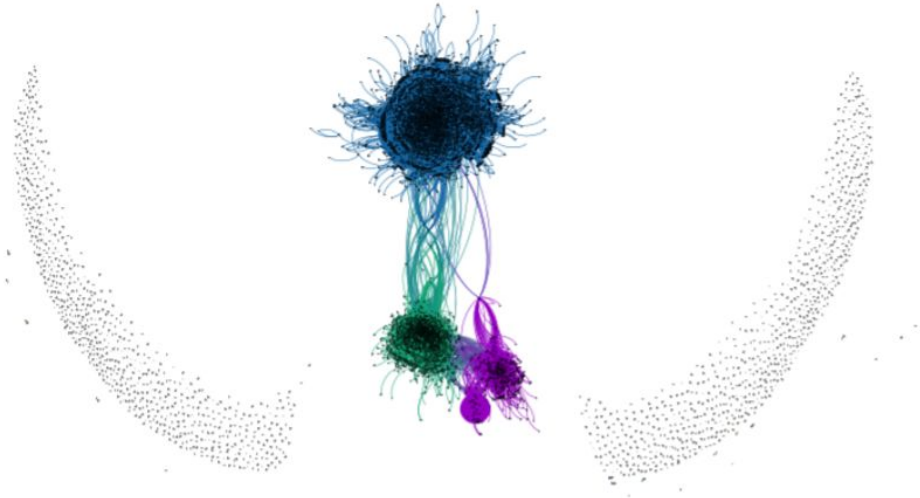


Fig. 1: *Interaction graph of all participants across all mailing lists explored in this study, rendered with [Gephi]. The large blue module is roughly the SciPy community. The green module is the Wikimedia community. The purple module is the OpenStreetMap community.*

History

- Developed to study open collaborative communities in 2015.
- In 2016, adapted to study **human rights advocacy** in IETF and ICANN
- Use and maintenance 2017-2019, used as a teaching and research tool.
- In 2020, Article 19 funds improvements to gender and affiliation detection, IETF datatracker and attendance ingest.
- In 2021, Article 19 sponsors BigBang Sprint at IETF 110.

University of Glasgow

UNIVERSITY OF AMSTERDAM

Berkeley

UNIVERSITY OF CALIFORNIA

DATACTIVE

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

Outcomes from IETF 110 Sprint: Software community

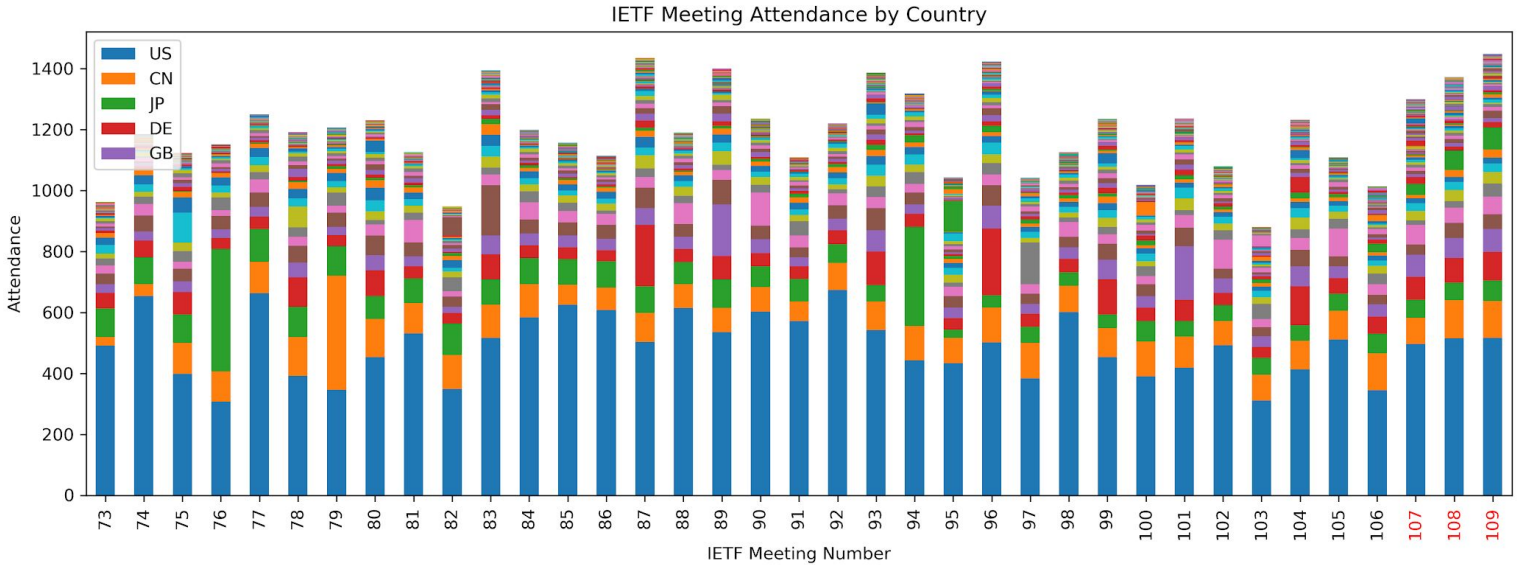
- *Growth.* New participants in the project!
- *Maintenance*. Updated installation instructions to keep up with dependencies.
- *Onboarding*. Produced instructional videos for installation and basic usage.
- *Debugging*. Debugged ingest issues around malformed data.
- *New data sources*. Work towards 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)
  - Well received and presented at the IETF Plenary!

- *Organizational involvement*. Building tools to better understand the involvement of organizations in IETF and other standards groups.
  - Received encouraging feedback and offer of help from Jari Arkko, the rfcstats etc. developer

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]

Meeting attendance 2019-2020

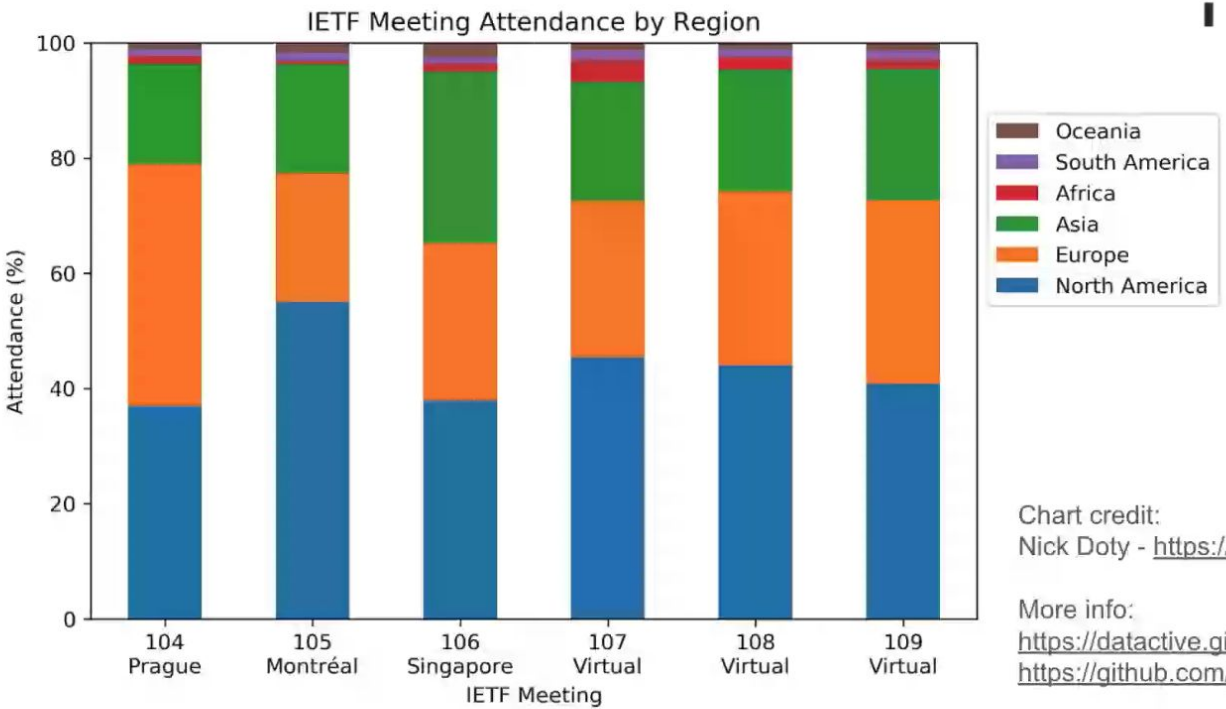
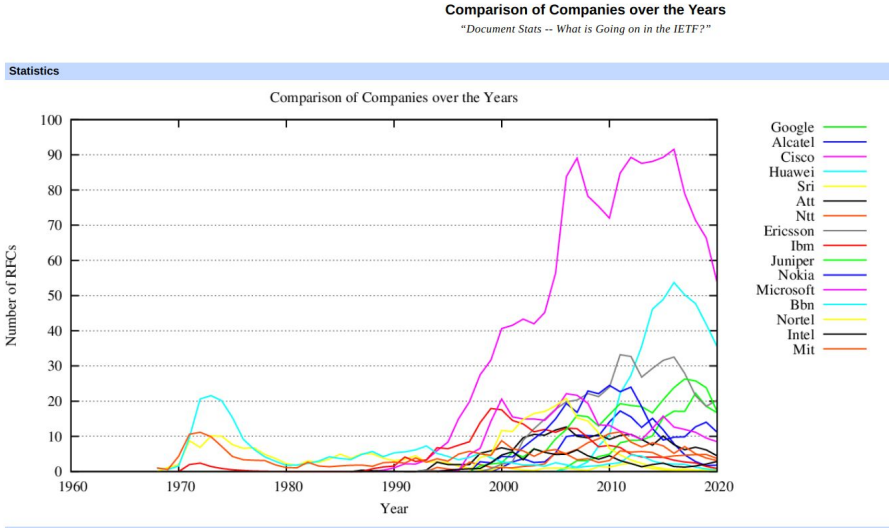


Chart credit:
Nick Doty - https://npdoty.name

More info:
https://datactive.github.io/bigbang/
https://github.com/glasgow-ipl/ietfdata

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.

BigBang @ IETF 110 in summary

- Captured enthusiasm about the project from new contributors and users
- Found broader IETF community supportive of the project
- There is an appetite for reflexive data science within IETF and openness to giving it a platform!

Future plans: Technical

- 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

Vision and Strategy:

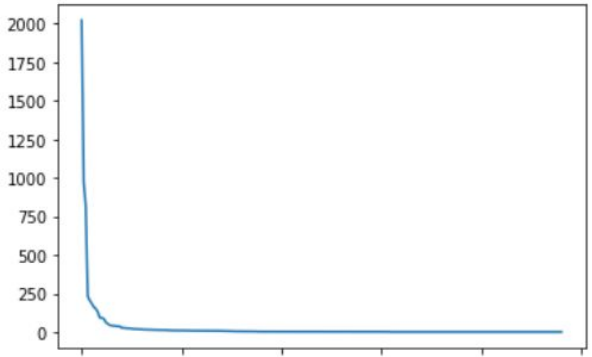
Data Science of Institutional Power

Strategic Plan for BigBang

- Theory of power
  - Networks-and-standards global governance circumventing “rule of law” protections of human rights. (Cohen, 2019)
  - Historically libertarian/meritocratic civil society institutions have become dominated by multinational corporations?
  - Activism and advocacy contests on a shifting terrain of technical foci
- Quantitative analysis (with BigBang)
  - Translate political research questions into quantitative analysis
  - Automated analysis for comprehensive view across institutions
- Actionable insights vis a vis human rights
  - Identifying “free agents”
  - Active monitoring of shifts in corporate involvement and attention

Example: Email domain analysis - httpbisa

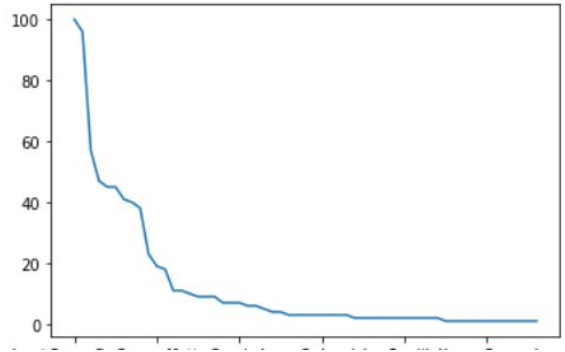
Top 10 Email Domains	Domain Entropy
google.com	3.140787
gmail.com	2.984459
hotmail.com	2.553237
oracle.com	2.058373
w3.org	1.758979
yahoo.com	1.744146
microsoft.com	1.699436
akamai.com	1.585919
ericsson.com	1.512764



Messages per email - gmail.com

Many differently affiliated individuals with gmail accounts are heavily involved -- at major differences in scale.

Area under curve indicating corporate strategy.



Messages per email - google.com

Example: Email domain analysis

- We assumed we could easily track corporate power's influence on IETF standards

- Our work was complicated by the fact that many prominent IETF participants present themselves as individuals.
  - Indeed, this is in accord with IETF policy: "participants engage in their individual capacity, not as company representatives. "

- We can use BigBang to better understand these individuals.
  - Does their authority increase as they change affiliation?
  - What happens if they get involved in human rights?

Email and other data sets

- Email is messy, but ubiquitous, communication medium used across standards bodies.
  - Provides a basis for comparison and way to track activity across fields.
- What about patents?
- Can email involvement in a technical area anticipate their patenting moves?

Vision: Active Monitoring for Advocates

- In-depth qualitative research and experience is essential to making sense of patterns in the data.
- But data signatures encoded in scripts can be used for *active monitoring* of standards-setting activity.
- We see BigBang as growing into a strategic interface for advocates engaging with standards bodies: who is gaining power? Who is making moves?
- Community interest in quantitative analytics also gives researchers a platform for presentation and intervention within the communities themseles.
    - E.g. IETF 110 Plenary talk.

Making it happen

- The BigBang community will be seeking continued support from NGOs to maintain *a regular presence at IETF hackathons.*

- Providing interesting insights for use in Plenary talks will increase the profile of the project and provide *vehicle for data-driven advocacy* within technical communities.

- Comparative analysis between different standards groups will inform scholarship and new approaches to human rights initiatives.

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

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