Towards a Better Understanding of Internet Protocol Standardization

An Analysis of the IETF Email Archives
A master thesis presentation

Cezary Radoslaw Jaskula

University of Oslo

June 18, 2021

The goal of this project

To develop methods and tools that will allow us to analyze the decision-making contained within the IETF email archives.

Why ?

The process that leads to the creation of these standards is not frequently undergoing a systematic analysis.

How ?

By parsing the email archives from their raw state and ingesting the data into a customized, semi-structured, full-text database building on the Apache Solr framework.

The IETF

- Internet Engineering Task Force
- Participants not members
- Working Groups
- Mailing lists
- Meetings three times a year
- Drafts - last only 185 days
- RFCs - the finished product
- The tools team

Toolchain

1. Email archives (mbox files)
2. Cleanarch
3. Parser
4. Solr
5. Queries (statistics)
6. Web scraper

The email archives

- Saved in mbox format
- Mbox format commonly used in unix distro
- Inherently flawed in the way messages are stored
- Solved by cleanarch

Cleanarch

- Script created to solve the mbox separator problem
- Uses Python 2
- Modified to run outside of the Mailman framework
- Uses "|" instead of ">"
- Used in this project to clean the mbox files
- Error = the presence of a default value in a mandatory field
    - From
    - To
    - Date
- Heavily reduced the error rate in the final database

Cleanarch results

Error combination	Before	After	Decrease
ALL	38809	20992	45.89%
From + Date + Dest	6934	443	93.61%
From + Date	29	22	24.13%
From + Dest	157	157	0%
Date + Dest	44	44	0%
From	89	89	0%
Date	12736	12736	0%
Dest	18820	7501	64.14%

Clean files produce 2893656 documents

File categories

- 3 categories
  - mbox
  - possibly mbox
  - not mbox
- Much higher error rate in Category 2
- Category 2 cleaned and added to the database
- Eliminated 1096 out of 11219 errors
- Clean category 2 files produce 106030 documents
- Content field made longer

Solr

- Core
- Schema
  - Defines fields and their datatype
  - Affects queries and results
  - Tokenizers
- Date field
- Fields with name and address pairs split into 3
  - Address
  - Name
  - Raw

Parser

- Python
- mailbox
  - Allowes to iterate over messages in a mbox file
- email.utils
  - Extract information from fields
  - Used to extract names and addresses
- Pysolr
  - Communication with the Solr instance
- Default value if extraction fails

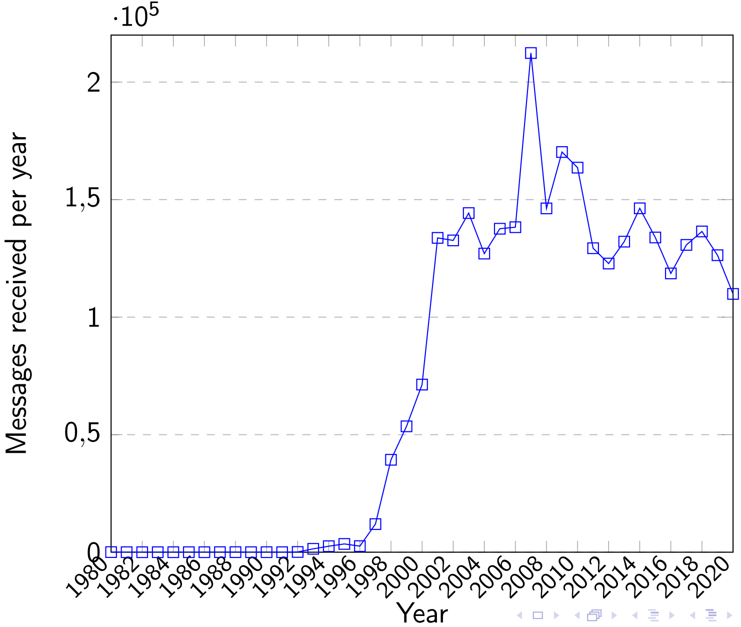
Faceting

- Query augmentation option
- Calculations done based on results set of query
- Counts tokens in a given field
- Used to calculate statistics
- Automatically sorted in decreasing order
- Example
  - query = *:*
  - facet.field = From-address

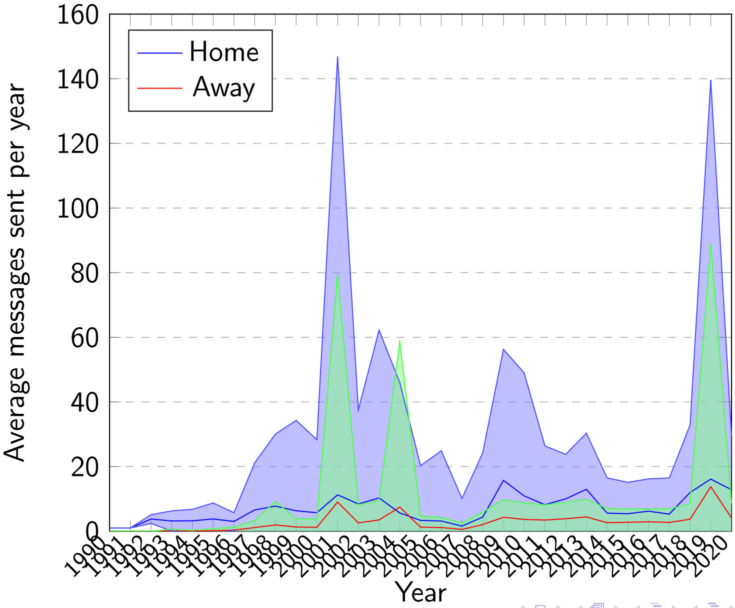
Ranking	address	Value
1	black_david@emc.com	14365
2	brian.e.carpenter@gmail.com	13491
3	marcelrf@bellsouth.net	11485
4	julian.reschke@gmx.de	10500
5	christer.holmberg@ericsson.com	10230
6	stephen.farrell@cs.tcd.ie	8567
7	martin.thomson@gmail.com	8522
8	cabo@tzi.org	7915
9	dromasca@avaya.com	7606
10	jari.arkko@piuha.net	7482
11	j.schoenwaelder@jacobs-university.de	7306
12	paul.hoffman@vpnc.org	7108
13	mcr+ietf@sandelman.ca	7101
14	mnot@mnot.net	7043
15	adrian@olddog.co.uk	6811

Ranking	address	Value
1	ietf	134157
2	i-d-announce	103544
3	quic-issues	54104
4	v6ops	44450
5	ips	40603
6	avt	40454
7	dmarc-report	40007
8	ipv6	38591
9	httpbisa	38133
10	mobileip	37475

Messages sent to all mailing lists per year



Crosstalk

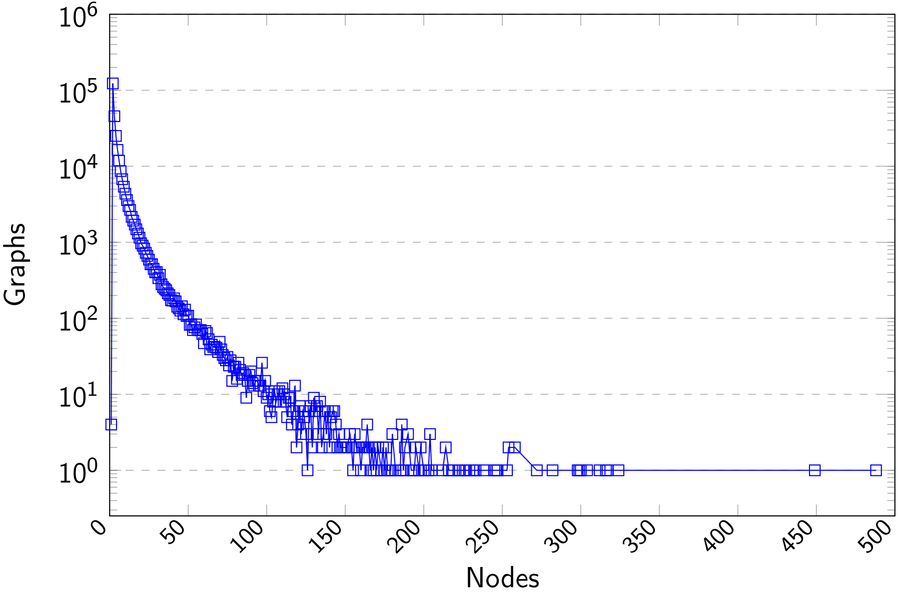


Home
Away
Average messages sent per year
Year
0
20
40
60
80
100
120
140
160
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

Conversation tracking

- In-reply-to-field
- Messages as nodes
- Parent - child relation
- Incorrect dates
- Relations mapped and represented as bidirectional graphs
- 2 cores
  - Nodes
  - Graphs
- Graphs are conversations

Resulting graphs



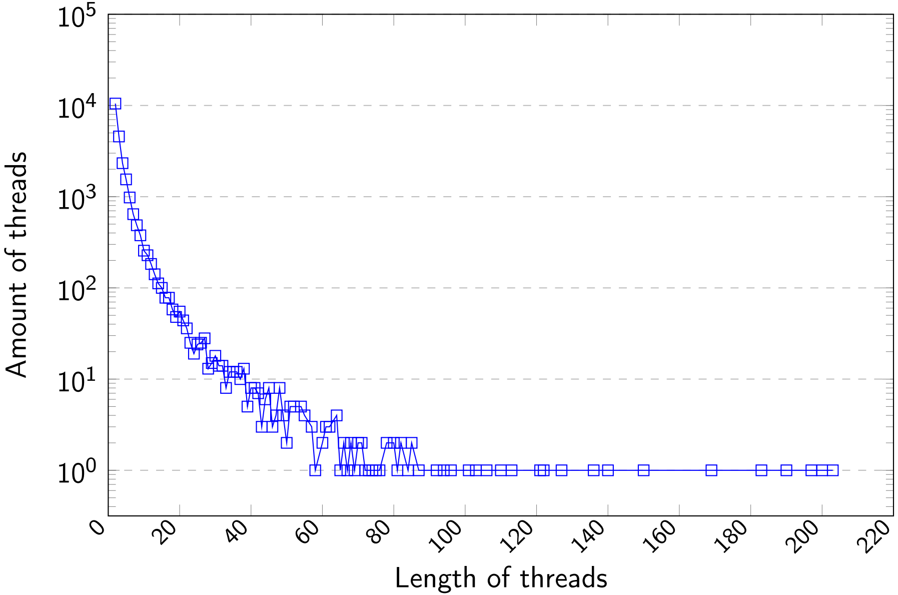
High level conversation tracking

1. Execute query on the “Subject” and “Content” fields
2. Identify the graphs the nodes in the results set belong to
3. Fetch graphs nodes from Solr
4. For each graph, set the nodes from the query results set as “targets”
5. For each target, try and find a path to any of the other targets
6. For each path found, save the nodes in it in a set
7. See if any paths have common nodes, if yes, merge them
8. Finally, for each result set, order the nodes by date

Tracking "Last call"

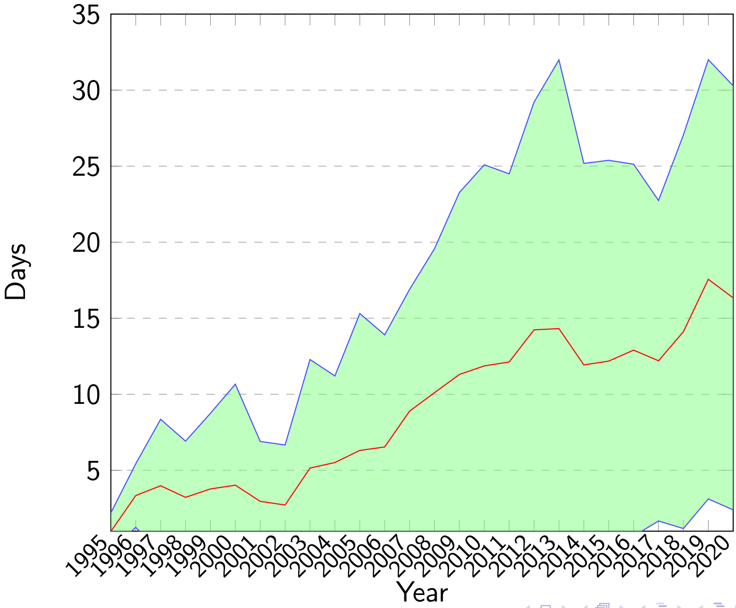
Amount of conversations found $= 23254$
Average length $= 4.567$
Standard deviation $= 6.931$

"Last call"

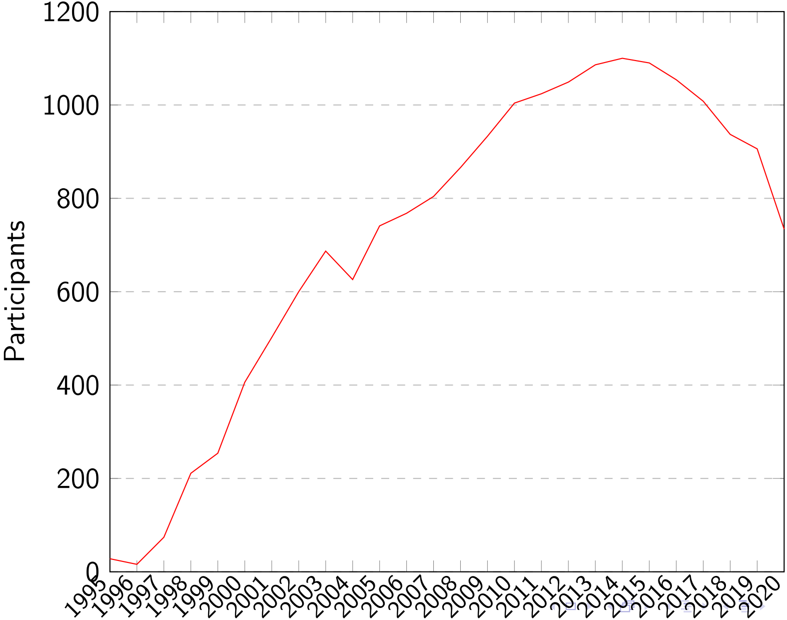


10^5
10^4
10^3
10^2
10^1
10^0
Amount of threads
0
20
40
60
80
100
120
140
160
180
200
220
Length of threads

"Last call"


Days
35
30
25
20
15
10
5
Year
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

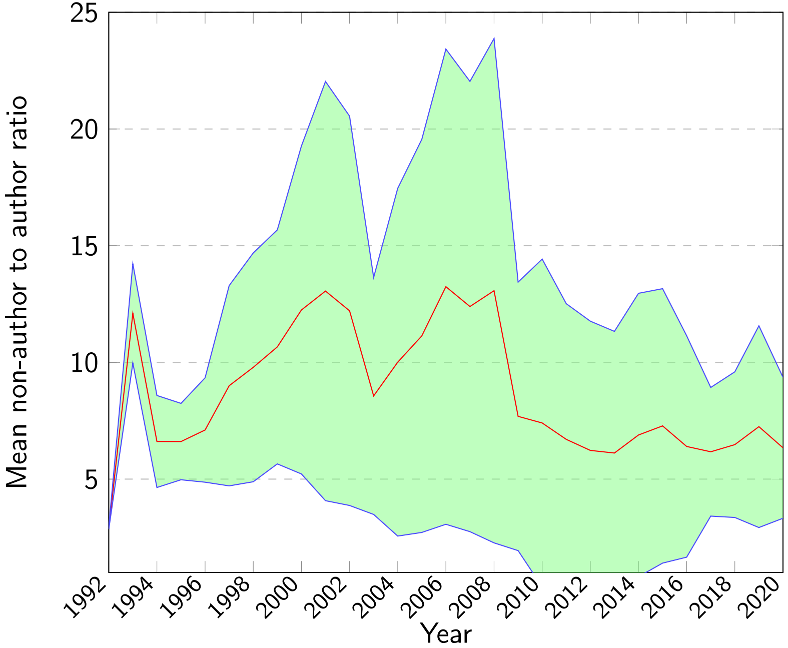
"Last call"


Participants
0
200
400
600
800
1000
1200
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

Connecting authors to threads

- LaTeX .bib file
- IETF datatracker
- RFC ids
- Web scraper
  - Beautifulsoup
- More addresses from Solr

Ranking	*Name*	*RFC count*
1	Russ Housley	96
2	Donald Eastlake	95
3	Keith McCloghrie	92
4	Henning Schulzrinne	90
5	Hannes Tschofenig	85
6	Yakov Rekhter	78
7	Jonathan Rosenberg	72
8	Adrian Farrel	71
9	Paul Hoffman	70
10	Gonzalo Camarillo	70
11	Marshall Rose	65
12	Fred Baker	65
13	Alexey Melnikov	60
14	John Klensin	59
15	Mohamed Boucadair	54


25
20
15
10
5
Mean non-author to author ratio
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Year

Summary

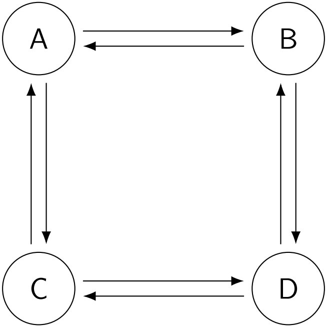
- Email archives parsed and transformed into a Solr compatible format
- Various statistics calculated
- Relations between messages have been mapped
- Basic thread tracking implemented and working as intended
- Authors and their addresses collected

Conclusion

- The IETF is an active forum with many users
- It is a good place to learn as people talk across mailing lists
- Both authors and non authors interact and contribute to discussion
- The "last call" threads are losing participants, and decreasing in length

The end

Thank you for your time


A
B
C
D