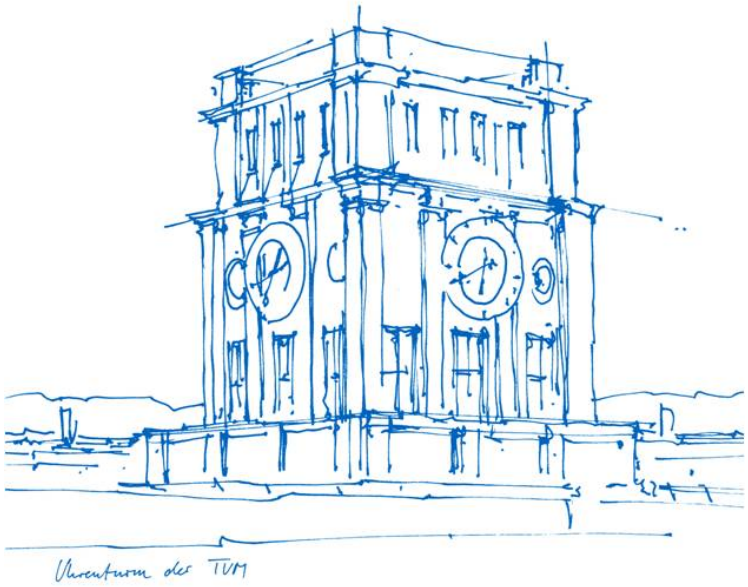
# Information Mining from Public Mailing Lists: A Case Study on IETF Mailing Lists

Heiko Niedermayer, Nikolai Schwellnus,
Daniel Raumer, Edwin Cordeiro, Georg Carle



# Overview

- IETF
- Dataset
- Person Deduplication
- Results

# Introduction to Dataset

- IETF (Internet Engineering Task Force)
  - Generates Internet standards
  - Open, operates via mailing lists, and meetings
  - Started in 1986
- Our dataset
  - Continuous collection of data from IETF servers (we clone the data)
  - Mailing lists
  - Documents
  - Meeting Information
  - …

# Problem of Identifying Individuals

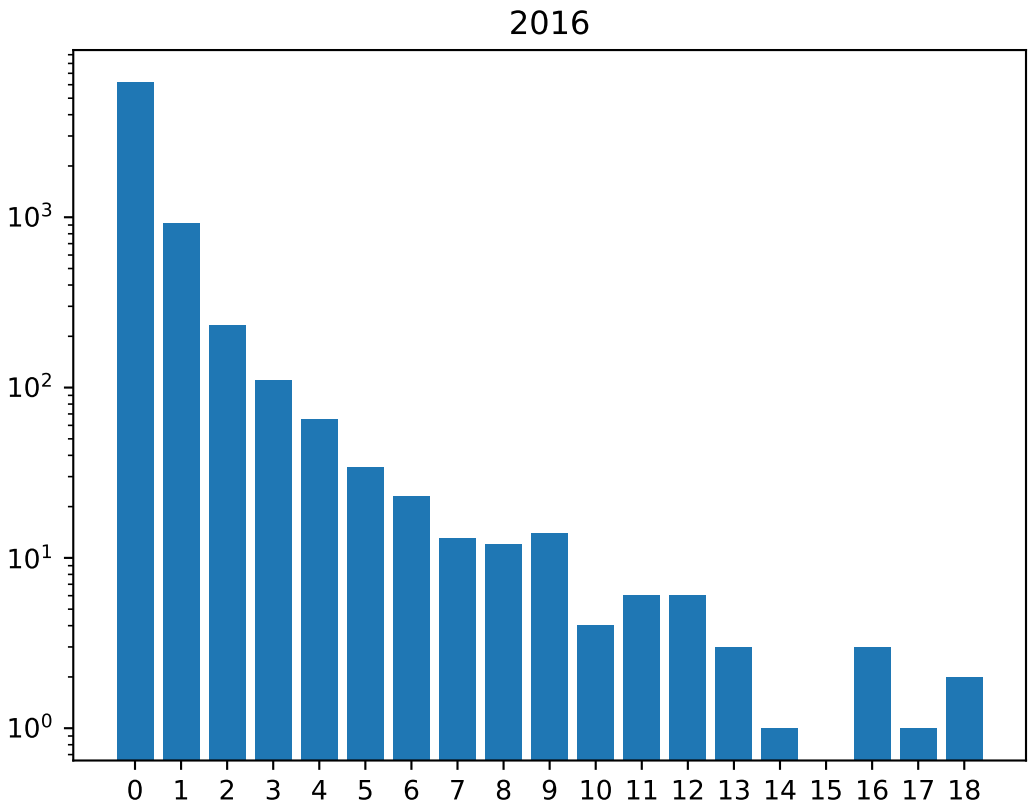
- Mailing List analysis has to deal with a variety of issues
  - Spam
  - Email addresses != Persons
- People change affiliation and addresses over time
- People may send from different accounts or in different roles
- Related Work suggests to heuristically determine
  - Which addresses belong together
  - Which name is the most meaningful name
- Limitation of Related Work
  - Smaller and less open lists were analyzed
  - Mostly English names vs international community (IETF)

# Deduplication of Individuals

- Heuristics
  - Same email, different names → merge
  - First and last name identical or in reversered order
  - Edit distance of full name small and name long enough
  - …
- Utilizing other sources
  - PGP PKI
    - Contains email addresses, names, keys
    - Cryptographically-signed
    - Multiple email addresses given → merge
    - Other email addresses not obtained via PGP can still belong to the person
  - In IETF case a lot of PGP keys found, less likely in other lists

# Results

# Participation Statistics




Now that we have better data about individuals, in how many lists are they active?

# How stable is individual posting behaviour over time?

**Table 1.** Mails $P[x_t \geq x_1 | x_{t+1} \geq x_2]$

| $x_1$ \ $x_2$ | 1 | 11 | 21 | 31 | 41 | 51 |
|---|---|---|---|---|---|---|
| 1 | 40.3 % | 12.2 % | 7.15 % | 4.83 % | 3.42 % | 2.52 % |
| 11 | 80.4 % | 48.3 % | 33.3 % | 24.0 % | 17.6 % | 13.2 % |
| 21 | 87.5 % | 63.5 % | 48.5 % | 37.6 % | 28.5 % | 22.1 % |
| 31 | 90.1 % | 70.9 % | 58.5 % | 48.0 % | 37.9 % | 30.1 % |
| 41 | 92.4 % | 76.9 % | 66.9 % | 57.5 % | 47.8 % | 38.7 % |
| 51 | 92.9 % | 80.7 % | 71.9 % | 63.5 % | 54.6 % | 45.4 % |

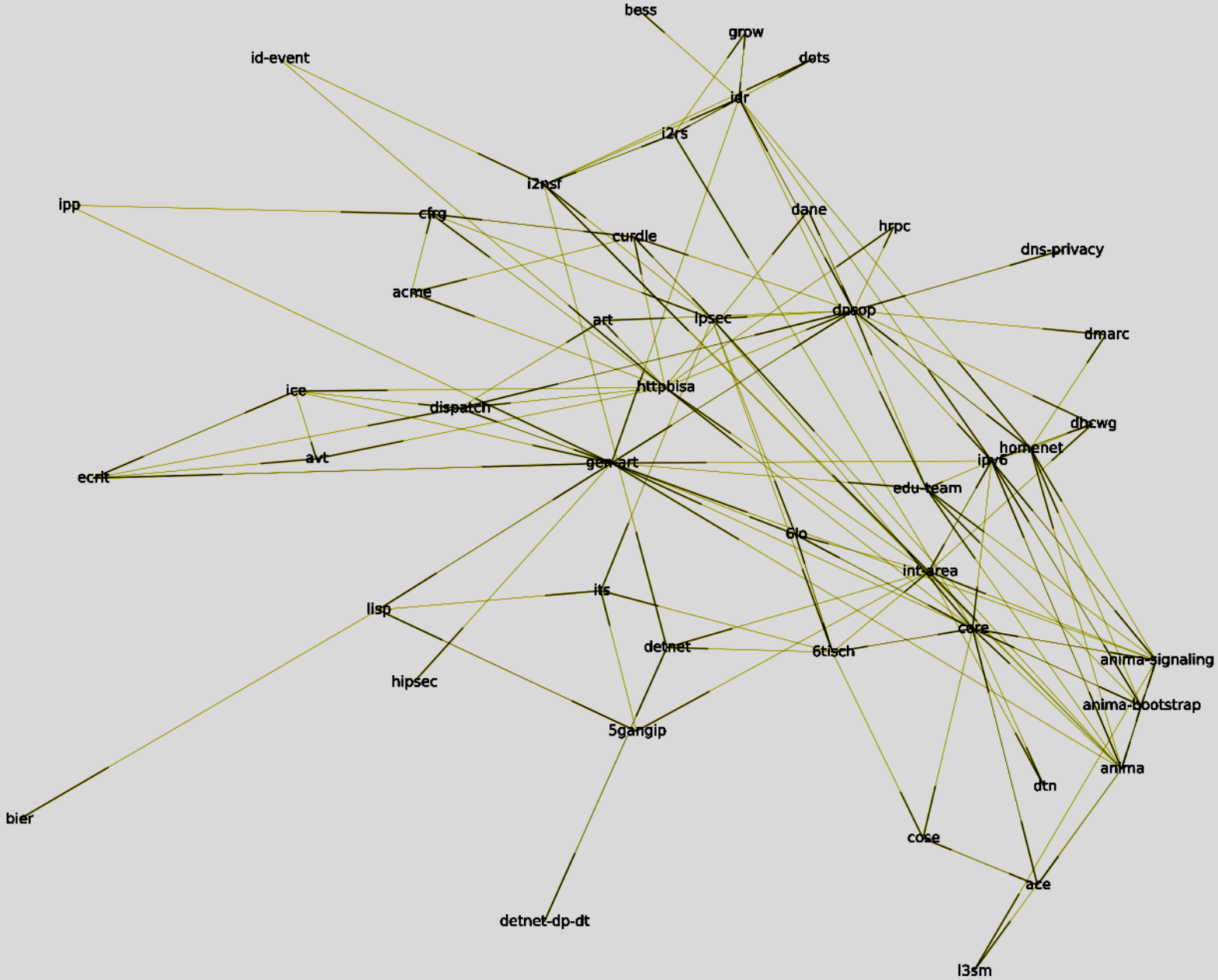
*If I post more than x1 messages this year,*
*I will post  more than x2 next year….*

# How stable is the posting behavior in number of groups?

**Table 2.** Lists $P[y_t \geq y_1 | y_{t+1} \geq y_2]$

| $y_1$ \ $y_2$ | 1 | 2 | 3 | 4 | 6 | 11 |
|---|---|---|---|---|---|---|
| 1 | 58.8 % | 33.7 % | 22.6 % | 15.8 % | 8.96 % | 2.95 % |
| 2 | 74.7 % | 58.0 % | 42.8 % | 31.5 % | 19.0 % | 6.45 % |
| 3 | 84.3 % | 73.3 % | 60.1 % | 47.6 % | 30.8 % | 10.9 % |
| 4 | 89.7 % | 82.2 % | 72.6 % | 61.9 % | 43.1 % | 16.6 % |
| 6 | 94.5 % | 91.3 % | 87.1 % | 80.3 % | 63.3 % | 28.9 % |
| 11 | 96.5 % | 94.5 % | 93.9 % | 92.3 % | 86.6 % | 62.4 % |

*If I post in more than y1 groups this year,*
*I will post in more than x2 next year….*




*Edge weight based on similarity of the two groups in terms of participants*

# Conclusions

- Internet Standardization
- Mailing Lists
- Person Deduplication
- Participation in Lists