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

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

Chair of Network Architectures and Services,
Department of Informatics, Technical University of Munich
{niedermh,schwelln,raumer,cordeiro,carle}@in.tum.de

Abstract. Public mailing lists, such as the mailing lists used by the IETF for Internet Standardization, can be used as big real world data set for analysis of social interactions. However, volatile participation and the usage of mail addresses as changeable pseudonyms constitute a challenge for data mining in these data. We conducted a case study of mailing list analysis wherein we address the consistent identification of a person with all of her contributions to be used as panel data. Based on the postings of individuals on different mailing lists, correlations between standardization areas in the IETF groups can be computed. Isolated and meshed standardization areas can be identified.

Keywords: mailing lists, identity deduplication, clustering, standardization

1 Introduction

Open mailing lists provide a vast area of open data for research on activities of the related groups. We have been studying the standardization efforts of the Internet Engineering Task Force (IETF) wherein open technical discussions lead to standardization of Internet technologies. Since these lists are open and can be accessed and joined via simple registration, a lot of different people from all over the world with different and changing backgrounds are represented. While the IETF and its contributing individuals are active for decades, the used mail addresses functioning as pseudonyms may change, e.g. due to the change of company affiliation. Thus, identifying an individual on the list simply by her e-mail address is not sufficient. Additionally, spam accounts and management accounts may also influence the data and they need to be addressed.

Our contribution is as follows: We applied and adapted the approach of Jensen et al.[6] for the deduplication of users. We analyze the outcome. Furthermore, having ensured that persons had been properly deduplicated, we present some results that can be generated from this data.

2 Related Work

At the 3rd International Conference on Internet Science, we presented our initial dataset of the publicly available information of the Internet Engineering Task Force (IETF)[10]. We presented a first analysis on this dataset and showed the influence of external occurrences like the Snowden leaks on security related standardization activities and vice versa the influence of internal IETF activities on the outside world in the social media service Twitter. This work is intended to update on the grown datasets and extended analysis.

Bettenburg et al.[1] used off-the-shelf algorithms to analyze data from mailing lists. They did not find any off-the-shelf solution for clustering of multiple identities by a single person but stated that sibling identities render social analysis useless. If not solved, persons appear multiple times in the data and her actions may be considered independent actions from different individuals.

The basic ideas for deduplication of persons on mailing list data is given by Bird et al. [2]. We based our approach on the work from Jensen et al. [6]. Maijuan et al. [9] present a metric to select the best name alias out of several grouped mail identities.

Jensen et al. [6] study the behaviour of new users (newbies) on a set of open source mailing lists. They found that these newbies got replies quickly. The replies are helpful most of the time and only a small percentage of rude replies occurs. Junior et al. [7] use neurolinguistic theory to obtain and mine information about developers on a mailing list. Chen et al. [4] generate social graphs on the basis of posts and replies on mailing lists. Toral Mar et al.[8] performed a factor analysis on an open source software mailing list. Other work centers around the case that e-mail data of an organisation is given and they try to infer the social network and status of individuals on the list, e.g. [11]. Due to different foci of the mailing lists – e.g. between software development focused and standardization focused lists, results are expected to differ.

3 The IETF Dataset

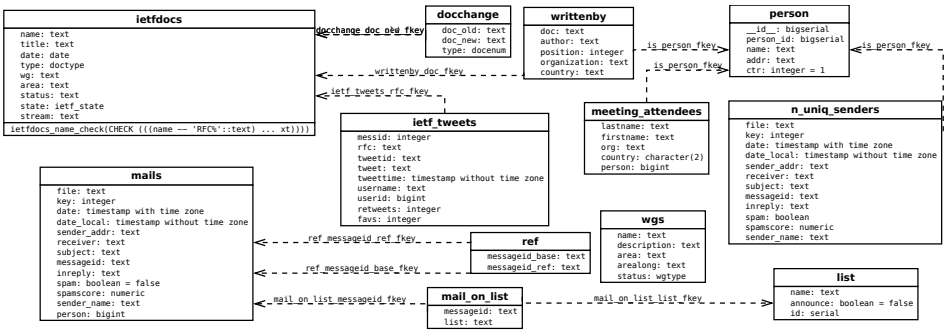

Fig. 1. Database schema

We already presented our initial dataset in previous work [10]. The current and extended database schema is shown in Figure 1. In the following, we focus on mailing lists. The data is generated from the parsed mailing list archives of all IETF mailing lists and the published drafts and RFC documents. We extract meta information for each mail. The archives include discussions and announcements as well as spam mails. We filter out spam mails on the basis of the provided spam scores. The database stores around 2 million mails posted on 984 mailing lists from around 20,000 mail addresses, which belong to 13,439 actual persons after deduplication as explain in the paper.

To understand inter-relations between different lists, we introduce unique persons that are identified by a surrogate id. Otherwise it is not possible to recognize when someone has sent mail from different accounts. In the following, we focus on lists where standardization-relevant technical discussions occur. We exclude management lists and meeting participation mailing lists from our analysis.

4 Identification and Deduplication of Individuals

Beside changes of mail addresses due to a change in an individual's company affiliation, we found that authors use different mail addresses on mailing lists than in the RFC documents. Other authors changed mailing addresses at certain times. These surrogate identities were not directly mappable to real individuals and presented a hurdle for further analysis. Similar problems have been reported in other cases of mailing list analysis. We follow the approach of Jensen et al.[6] with slight modifications.

4.1 Algorithm

When we fill our database with entries for a new person, they have a name given from the e-mail header and an e-mail address. If these values differ from previous entries, it will naturally be a new entry. Then the deduplication algorithm tries to find non-matching previous entries that are similar enough to merge new and old entries into one. In the database they will both exist, but share the same person identifier.

In the preprocessing, we have to normalize the name representation. This means to remove whitespace, dots, and titles in the name. Like related work we split the name into first name, middle name, and last name. Additional normalization would include normalization of special characters not found in English. We also blacklist spam names and spam e-mail addresses. We also want to exclude management accounts which might get merged due to similar names and might not contribute to the standardization discussions that we would want to analyze in our subsequent work.

We merge two persons if:

- their e-mail addresses are identical, even if names differ.

- first and last name are identical or in reversed order
- first and last name are contained in the other e-mail address or name
- the full names are similar enough. We use the Levenshtein edit distance to anticipate minor changes in the writing of names. The decision is then based on the metric $S = 1 - \frac{Distance(name1,name2)}{max(len(name1),len(name2))}$ and we merge if $S > 0.85$
- if the e-mail address before @ is equal, but shorter than 6 characters, yet the full name similarity $S > 0.75$

In addition to name and e-mail address similarity, we propose to use additional external input. Most promising for mailing lists are PGP keys and their signatures provided by keyservers given enough security-aware persons are participating in the mailing lists.

4.2 Using information from PGP keyservers

PGP key servers provide a ground truth for clustering e-mail addresses together as the owners of the e-mail address put them together in a PGP signature associating all of them with their PGP key. This information shows that two mail addresses belong to the same person if they share the same PGP key. It does not show, however, that two e-mail addresses are not from the same person. We can use this to improve our algorithm and make better decisions when PGP is present. If we have two entries, but both e-mail addresses use the same PGP key, we will merge the two persons into one person.

PGP entries also include the name. So, the PGP entries may give the most preferable name for the person. In our study, for 55.1 % of our person entries we could find PGP keys via public key servers.

5 Persons and Groups

5.1 Statistics about the deduplication

Overall, there are around 13,000 persons in the dataset after the deduplication. About 10,000 did not need deduplication, around 3,000 are the result of merging persons that the deduplication considered to be the same person. In normal cases the number of persons merged into one is at most slightly above 10. However, there are some larger cases in our data set. The largest one with 86 persons is from merging spam accounts that made it through the spam defense. Another issue is that persons with long identical first names might generate a situation where they get falsely merged. 46 persons got merged from more than 10 persons each, 264 from more than 5 persons each. Considering the overall number of over 13,000 persons this is a small number below 2 %, but there is an increased chance among these 264 that they may have been generated from a false merge.

In our subsequent studies, in particular, the spam accounts are blacklisted and will not affect the outcome. Merging two low-profile persons will have little impact, merging a high-profile and a low-profile one as well.

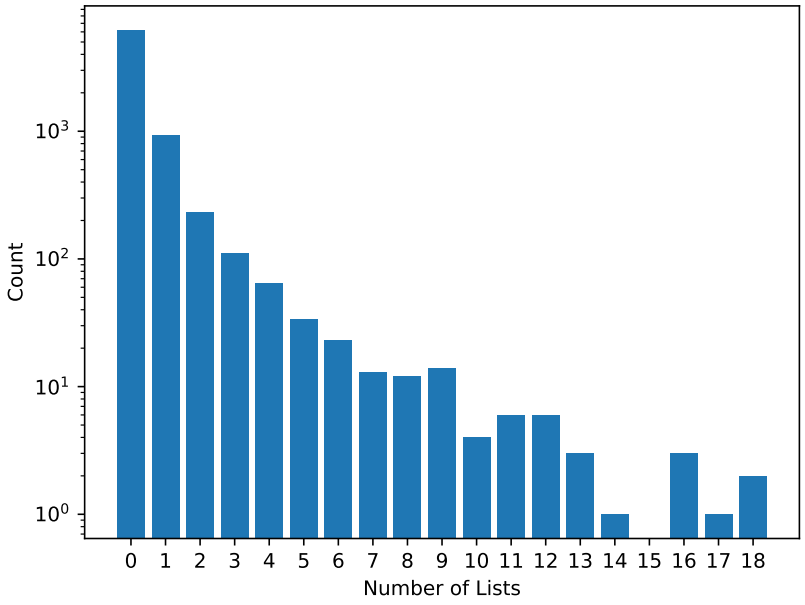

Fig. 2. Number of mailing lists a person posted in 2016

5.2 In how many groups do people post?

Figure 2 shows the histogram for 2016. Most people have not been active in 2016 since our data set covers mails from as early as the year 2000. Participation has changed over time. 925 posted in only one list, 233 in two lists, 111 in three lists, and 65 in four lists. The maximum is 55 lists.

One may wonder if one can really speak of participation in a list if only few messages over a whole year were sent. If we only count lists where a person has at least 10 posts, the situation is as follows: only 368 people make it in at least one list, 90 in two lists, 25 in three lists, and 12 in four lists. The maximum is 11 lists.

5.3 Who is posting in many different groups?

While most people only post in one or two mailing lists, there are a few individuals that post in many lists. In 2016, an individual posted in 55 lists (32 with more than 3 message), one in 39 (25 with three messages, 11 with at least 10 messages), and one in 28 (19 with at least 3 messages). All of them had occupied area director positions at the time and were responsible for a larger set of working groups which each has at least one list. The numbers drop significantly when we only count for lists with a certain amount of messages. This shows that

although these persons post on a broad range of lists, their predominant activity in discussion only focuses on a subset of these lists.

5.4 How stable is the individual posting behavior 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 %

A person can either post in a group or not post in a group. We are now interested in the change from one year to another. We used a subset of our whole data that covers the last 10 years from 2007 to 2016.

In Table 1, each row refers to the number of posts in the year before. The first row means that at least one message was sent in the group by the person. The table gives the percentage of such users that one year later still post at least one message in that group (first column), 11 or more messages in that group, and so forth. In the next row, the user has posted 11 or more messages in the previous year and so forth.

Those who only posted at least one message in a group will in 40.3 % of the cases post atleast one message in the following year. A majority will not. However, those who post regularly, will also post a lot of messages in the subsequent years. Even for the ones with a lot of messages (over 50) in a group, 7 % will not post in the next year in this group. The likelihood that one will continue to post with a similar high rate, however, is below 50 %.

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

Table 2 shows statistics for the number of groups a person posts messages in. Those who posted in at least one group also post in at least one group in the subsequent year with 58.8 %. The percentage here is higher than in the previous section because people may switch to another group and, thus, not return to a group, but still return to another group. So, taking part in the IETF overall is more stable than taking part in an individual group.

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 %

6 Inter-group Analysis

After clustering mail addresses of the same individuals, we demonstrate how the dataset can be used for the analysis of group relations. Therefore, we build a social graph which is a common way to analyze human interaction (cf. [3]). We define two groups as related when individuals posting in one group are also posting in the other group. Based on this relationship, we build a directed graph wherein each mailing list represents a node. The edge weights are computed as one minus the reciprocal of the sum of mails on the other mailing list from each poster on the origin mailing list in the fifth potency.

Figure 3 shows a spring force graph representation plotted with the Python NetworkX library. The thickness of the graph at the origin side of the edge represents the edge weight. We filtered working groups with less than 100 posts in 2016. And only added edges with an edge weight of at least 0.1.

Although not relying on domain specific knowledge, the graph shows relations that are expected due to the relation of topic of the working groups or the IETF internal structure. The General Area Review Team (Gen-ART) mailing list takes a central position. On this list, leaders of working groups discuss and review documents with an IETF wide perspective. Also some non-general area related working groups take central positions like the Domain Name System Operations (dnsop) or Constrained RESTful Environments (core) working group. This reflects the strategical importance of these activities for other working groups.

The graph also shows subgroups with stronger intra-group links. Mailing lists like anima-signaling, anima-bootstrap and anima (all concerned with facets of autonomic networking) are positioned close to each other. Another subgroup is constituted by the security related working groups dane, ipsec, dnsop, and httpbisa; a third example is i2nsf, i2rs, and idr which are concerned with global routing. The graph also allows to identify weak ties. In analysis of social graphs, weak ties received high attention [5] due to their potential high effects that they can have on the groups connected via them. The groups bier, bess, hipsec, detnet-dp-dt, and dns-privacy only have one link to the other groups. This supposes that the connection has a strong character denoting high influence.

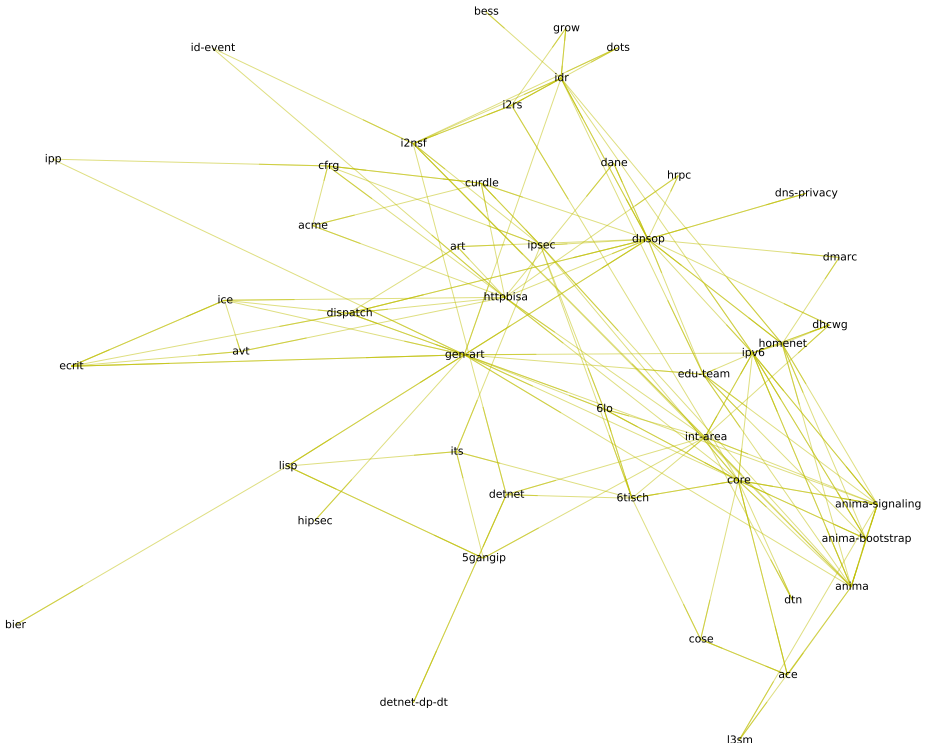

Fig. 3. Spring force graph of working groups

A closer look at the edges reveals another aspect of the relations: only some of them are symmetric (e.g. dnsop and dns-privacy). The relation between ipv6 and dnsop is asymmetric: while ipv6 members are also active in discussions in dnsop this statement cannot be made for members of dnsop that are active in ipv6.

7 Conclusions

We have presented the mailing list aspects of our IETF data set. We motivated the need for deduplication of persons and presented an adapted approach used in related work. We propose to enhance the mechanism using external sources like PGP keys from public keyservers. In a tech-savvy group like the IETF chances are high that such keys exist and improve the data set. Subsequently, we presented some statistics on how persons contribute to individual lists and to multiple lists. Finally, we provide a first visual inter-group analysis on the basis of overlap of group participants.

Indeed, contributors to IETF mailing lists are tech-savvy people. PGP keys are less likely to exist in less technical areas. Our data set covers a long period of

time. Thus, changes in affiliation and email address are likely to occur. Furthermore, the users are spread around the world and issues with non-English names play a significant role. This specificity of the studied mailing lists surely has to be considered when transfering our results to other areas.

References

1. Bettenburg, N., Shihab, E., Hassan, A.E.: An empirical study on the risks of using off-the-shelf techniques for processing mailing list data. In: 2009 IEEE International Conference on Software Maintenance, Edmonton, Alberta, Canada. pp. 539–542 (Sept 2009)
2. Bird, C., Gourley, A., Devanbu, P.T., Gertz, M., Swaminathan, A.: Mining email social networks. In: Proceedings of the 2006 International Workshop on Mining Software Repositories, MSR 2006, Shanghai, China. pp. 137–143 (May 2006)
3. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. science 323(5916), 892–895 (2009)
4. Chen, H., Shen, H., Xiong, J., Tan, S., Cheng, X.: Social network structure behind the mailing lists: ICT-IIIS at TREC 2006 expert finding track. In: Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, USA (Nov 2006)
5. Granovetter, M.S.: The strength of weak ties. American journal of sociology 78(6), 1360–1380 (1973)
6. Jensen, C., King, S., Kuechler, V.: Joining free/open source software communities: An analysis of newbies' first interactions on project mailing lists. In: 2011 44th Hawaii International Conference on System Sciences. pp. 1–10 (Jan 2011)
7. Junior, M.C., Mendonca, M., Farias, M., Henrique, P.: Oss developers context-specific preferred representational systems: A initial neurolinguistic text analysis of the apache mailing list. In: 7th IEEE Working Conference on Mining Software Repositories, MSR 2010. pp. 126–129 (May 2010)
8. Marín, S.L.T., Martínez-Torres, M.R., Barrero, F.: Modelling mailing list behaviour in open source projects: the case of ARM embedded linux. J. UCS 15(3), 648–664 (2009)
9. Meijuan, Y., Qingxian, W., Shuming, C., Xiaonan, L., Xiangyang, L.: Ranking the authority of name aliases for email users. In: 2011 Third International Conference on Multimedia Information Networking and Security. pp. 425–430 (Nov 2011)
10. Niedermayer, H., Raumer, D., Schwellnus, N., Cordeiro, E., Carle, G.: An Analysis of IETF Activities Using Mailing Lists and Social Media . In: Proceedings of the third international conference on Internet Science, INSCI2016. Florence, Italy (Sep 2016)
11. Yoo, S., Yang, Y., Lin, F., Moon, I.C.: Mining social networks for personalized email prioritization. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 967–976. KDD '09, ACM, New York, NY, USA (Jun 2009)