Identifying temporal trends in IETF participation

Wes Hardaker, Genevieve Bartlett
hardaker@isi.edu, bartlett@isi.edu
USC/ISI

September 29, 2021

1 Introduction

Researchers at USC/ISI have begun performing some large scale communication analysis using data from datasets like those from the IETF RFC, Internet-Draft and E-Mail archives. Although this work is very much work-in-progress, below we show some preliminary results in analyzing datasets that show the fruitfulness of our larger plans. We specifically look at labeled datasets that contain markings for organizations, countries, and authorship.

2 Example Analyses

We offer three examples of the type of analysis that can be performed by studying the temporal and authorship analysis aspects of IETF and other data. First, we examine the history of RFCs published by the top organizations over time §2.1. We then also compare analysis of looking at international participation communication graphs in an initial analysis of two large IETF mailing lists §2.2. Finally, we show a metric-learning based mechanism that may be able to identify common authorship components across sets of documents to help alleviate some the struggles with authorships in RFCs.

2.1 Grouping by organizational type

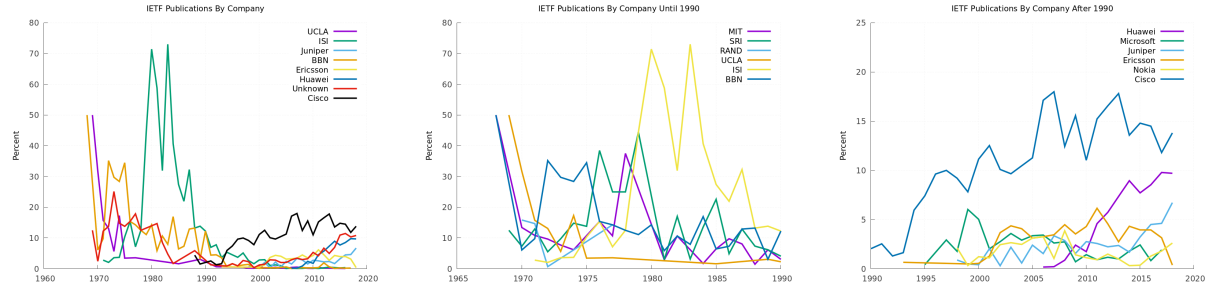
Looking at the longitudinal dataset of RFCs and their authorships, we find authors being employed by sets of organizations and companies. Using the analysis data created by Jari Arkko [1] and the IETF itself [3] and analyzing this data over time we study the rise and fall of individual organizations, countries and organizational types. We find that the longitudinal trends produces more significant insights than by looking at the collection of dataset as a whole.

If you look at the top few RFC authoring organizations over the entire IETF history (Figure 1a), you'll note a mix of types of affiliations. You may quickly note that some organizations have stopped participating (or participate significantly less) at points, while new organizations have popped up. However, if you break the top producing organization graphs in two, with periods before and after 1990 you can quickly see a massive change in the general participation within the IETF. Prior to 1990 (Figure 1b), the top contributing organizations were academically or research-based organizations like universities and labs. Post 1990 (Figure 1c), however, the top contributing organizations were predominantly for-profit companies, signifying both the increase in Internet commercialization and decrease in research and academic involvement.

Note: we have similar analysis at a country level, but leave out of the paper for brevity.

2.2 Grouping by country E-mail participation

To gauge a sense of international communication paths, the E-Mail archives [2] of the IETF offer a wealth of information. To study to where and from where E-Mail is being sent through the IETF's mailing lists, ISI researchers took a small single, recent year's sample of the traffic through the high volume *ietf* and *oauth*



(a) Looking at the top organizations over the IETF history

(b) Looking at the top organizations until 1990

(c) Looking at the top organizations after 1990

Figure 1: Top few organizations producing RFCs

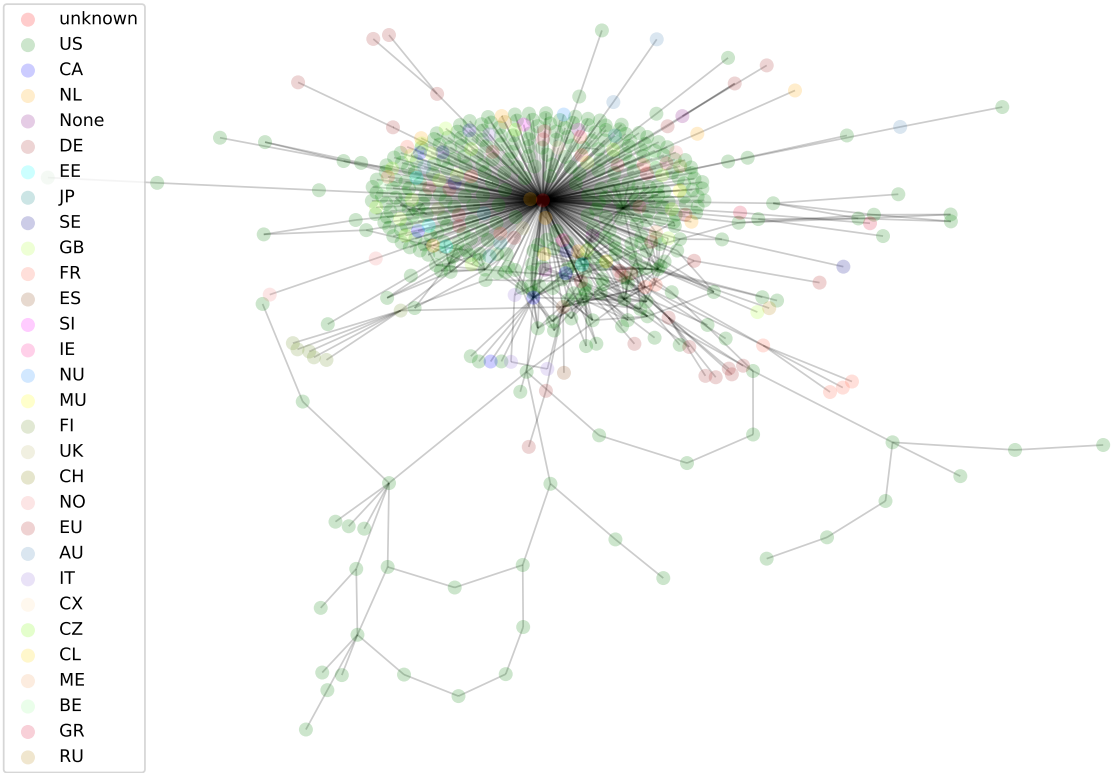


Figure 2: Example communication graph with country code labels constructed from a year of traffic through the high volume *ietf* and *oauth* mailing lists.

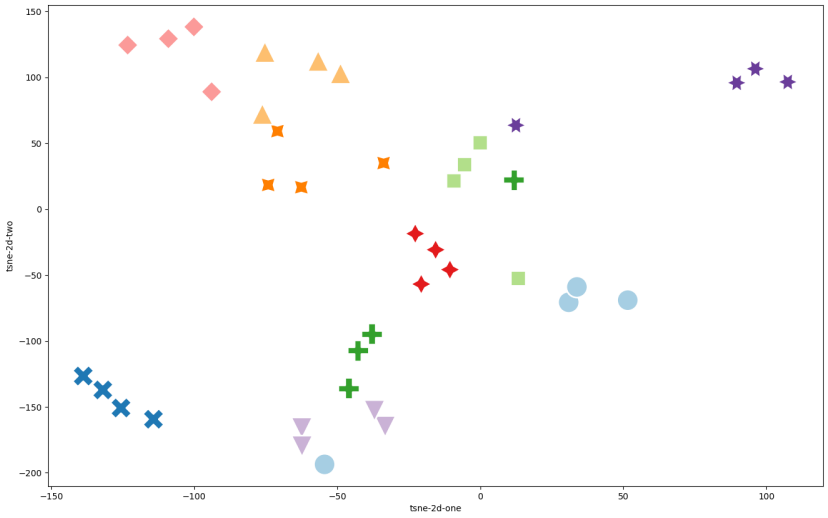


Figure 3: 2D t-SNE representation of author clustering for Avocado IT Emails Using Reddit Comments for Training. Each author identified by a unique color/shape.

mailing lists. From the headers we construct a graph network capturing the communication paths, showing the mailing list server at the center and participating organizations and mail servers at the outskirts Figure 2.

This work is only preliminary in nature, but we hope to expand it in order to identify the International footprint of communication within the IETF. Most importantly, we hope to study participation trends within IETF communication over time to identify temporal diversity characteristics in international participation.

2.3 Identifying authorship with stylistic clustering

To help breakdown authorship of individuals as they traverse organizations and utilize multiple E-Mail addresses, as well as potentially identify authorship of individual RFC sections, we plan to use some new stylistic machine learning techniques. We have not started applying these techniques to the IETF mailing list, internet-draft and RFC datasets yet, but we have demonstrated its ability to act on E-Mail archives from the Avacado IT dataset [4].

We approach the challenge of linking multiple samples of text based on authorship as a metric learning task. Our resulting learned featurization minimizes distance between feature vectors from the same author and maximizes the distance between differing authors. We train on broad and diverse sets of examples from many authors talking on multiple topics over time to deemphasize style features which change across time, social group and topics discussed.

Figure 3 depicts a 2D t-SNE representation of our learned features over 10 unique E-Mail authors—each author is represented by a unique color/shape from the Avacado IT dataset [4]. This dataset is a public archive of E-Mails from a defunct information technology company from the early 2000s with ground-truth for message authorship. Each point represents 4–10 emails from the same address. We clustered E-Mail messages covering a range of topics for each author, such as technical discussions, billing questions and social events, with each author discussing multiple topics between different social groups. Largely, the E-Mails from each author cluster closely regardless of topic, with a greater distance between each author's cluster and other authors. This indicates we can obtain fairly automatic authorship results based on E-Mail and RFC text samples, despite both sources having fuzzy authorship labeling.

3 Conclusions and future work

Although our work is very early in nature, we hope to extensively expand our research and develop techniques that apply to both the IETF specifically and that will generalize the tools and techniques to apply to other large datasets available around the Internet. The authors are currently applying for funding grants to support this future effort at a larger scale.

References

[1] Jari Arkko. Ietf statistics. `https://www.arkko.com/tools/stats.html`, 2021 09.

[2] IETF. Ietf email lists. `https://www.ietf.org/how/lists/`.

[3] IETF. Draft/rfc statistics. `https://datatracker.ietf.org/stats/document/author/documents/`, 2021 09.

[4] Douglas Oard, William Webber, David Kirsch, and Sergey Golitsynskiy. Avocado research email collection. *Philadelphia: Linguistic Data Consortium*, 2015.