Internet Complexity Through Five Decades of RFCs

ABSTRACT

Operation of the Internet requires interoperability between networks, systems, and applications, as well as cooperation among a growing number of stakeholders. The RFC series plays a critical role in supporting this cooperation and interoperability. As the series marks its 50th anniversary, we measure the shifts and trends that have emerged, including the rise and fall of large Internet players, and the shift in relevance between industry and academia. In addition, we show that the Internet's growth and maturity has given rise to a longer standardisation process, and provide initial insight into the growing complexity.

1 INTRODUCTION

The behaviour of the Internet is defined by a bewildering array of protocols. For more than 50 years, from the earliest experimentation with the ARPANET to the modern Internet, these protocols have been described in the RFC series of documents[1] (https://www.rfc-editor.org/). The RFC Editor publishes documents from the Internet Engineering Task Force (IETF), and other organisations, that bring together a diverse set of stakeholders to debate and standardise the protocols and processes that define the network. The RFC series therefore provides a unique historical record of the development of the Internet, and of how the process by which the Internet is developed has evolved.

Despite the importance of the RFC series, this diverse body of data has so far seen little analytical attention. We argue that by studying the RFC series and the Internet standards process, we can begin to understand how the complexity of both the network, and of the standards and processes that define the network, have grown over time. For example, this can not only reveal the social process that has driven decades of protocol design but also potentially shed light on major bottlenecks that could enable future streamlining of the standardisation process.

In this paper, we perform the first measurement study of RFCs to shed light on 50 years of IETF activities. To this end, we have collected a dataset comprising 8643 RFCs, from 5548 authors, working at 2175 institutions. Using this, we perform a bibliometric analysis that shows how the scope and complexity of the protocols, and the standards development process, have evolved. Our analysis is driven by two perspectives. We start by investigating the generation of RFC documents, particularly in terms of their spread across categories (§3). We then study the patterns of authorship and collaboration (§4). Our analysis reveals how RFC data can be exploited to shed light on network protocol standardisation activities, as well as delineating a set of bibliometric methodologies that can be used to achieve this. We also provide preliminary evidence regarding the complexity of the underlying social process that drives standardisation. To the best of our knowledge, this is the first paper studying the social dimension of the development of Internet protocols.

2 BACKGROUND & DATASET

The RFC series started in 1969 to organise work-in-progress notes of the community creating the ARPANET. As the ARPANET grew and became the modern Internet, the RFC series also evolved, becoming the primary publication series for technical standards developed by the IETF, and other documents relating to the development of the Internet [5].

RFCs were originally relatively informal working notes, but the series has become more formalised and structured over time. It currently comprises four active *publication streams*, one for each of the organisations that produce RFCs: the IETF stream, the Internet Research Task Force (IRTF) stream, the Internet Architecture Board (IAB) stream, and the Independent Submission stream. The IETF is an open standards forum that develops technical specifications and standards for the Internet. The IRTF is an associated venue that conducts longer-term research. Both are organised as activities under the IAB. Independent submissions provide a non-standards route to RFC publication, with a separate review process. The publication streams coordinate closely to avoid publishing conflicting material. The RFC Editor provides archival publication services for the streams. RFCs published prior to the formalisation of publication streams comprise the Legacy stream.

Each RFC has a *category*: standards track, best current practice, experimental, informational, or historic. Some RFCs prior to the formalisation of categories have 'unknown' category. Not all RFCs are standards, and the standards track is subdivided into proposed, draft, and full standards depending on maturity. Only the IETF stream can publish RFCs on the standards track, and only the IETF and IAB streams can publish best current practice RFCs. Experimental and informational RFCs can be published by any stream, and RFCs from any stream can be categorised as historic [7, 9, 10]. Each stream has its own procedures and review processes. Early RFCs were published with only limited review, but all RFCs now undergo extensive peer review prior to publication [1].

[1] The name RFC used to stand for "Request for Comments" and is a hold-over from the early development of the ARPANET.

To gather data about RFCs we rely on the RFC Editor, who maintains the master archive of the RFC documents, along with an index providing metadata about the RFCs, their authors, and their inter-relationships. From this index, we extract the following data: title, publication date, author names and affiliations, relationship with other RFCs, category, and publication stream. For RFCs published on the IETF stream, we also extract the name of the Working Group that developed the RFC, and the area of the IETF in which that working group was chartered. The IETF is organised into working groups with a focus on a particular topic that belongs to a more general technical area.

We augment this dataset with information from the IETF datatracker (https://datatracker.ietf.org). The datatracker is the main administrative database used by the IETF and IRTF. It records the status of work-in-progress *Internet-Drafts* that eventually become RFCs, as well as information about participants, working groups, meetings, etc. This information can be used to track the development of documents before they're published as RFCs. The datatracker was introduced in the early 2000s, and contains little historical data prior to that, but holds comprehensive information about more recent IETF activities. We use information from the datatracker in Section 3.2 to make an initial attempt at quantifying the impact of growing complexity within the standardisation process. The datatracker provides a REST API, that we access with an open source library we developed (which we will make publicly available, but anonymise due to double blind).

In total, we collect information on 5548 authors, from 2175 organisations, from 8643 RFCs for the period 1969–2019.

3 DOCUMENTS

We begin by looking at the documents published in the RFC series, how they fit into the different publication categories and areas, and their inter-relationships.

Figure 1a shows the cumulative number of publications in each category over time. Three phases can be seen in the development of the RFC series. First, in the period 1969-1974, RFCs are published at a rapid rate during the initial development of the ARPANET. Second, growth in the period 1974–1985 slows, as the relatively small community gains experience with using the network and develops the core set of applications and protocols. Finally, with the creation of the IETF and the introduction of the National Science Foundation Network (NSFNET) in 1986, the community and the number of RFCs published starts to expand rapidly. This is further driven by the opening of the network to commercial and public use in the mid-1990s, and continues to this day.

RFCs initially comprise a homogeneous series, as the series captures research notes and ideas during the initial development of the network. Categories and the three-stage standards track (*i.e.*, proposed, draft, and full standard) are introduced shortly after the formation of the IETF in the mid-1980s, as the network begins to see production use. New RFCs since that time have been broadly split between standards track and informational categories, with much smaller numbers of best current practice and experimental RFCs, and a steady trickle of older documents (about 318) being classified as historic as they become obsolete.

There is a steady growth in the number of standards track RFCs published, but it is clear that the three-stage standards track was not successful. Very few RFCs are advanced from proposed standard, to draft standard, to full standard. Recognising this, the IETF abolished the draft standard maturity level in 2011, leaving only a two-stage standards process [8]. This change also does not seem to have been a success. The Internet still runs primarily on proposed standards and few RFCs advance to full standard.

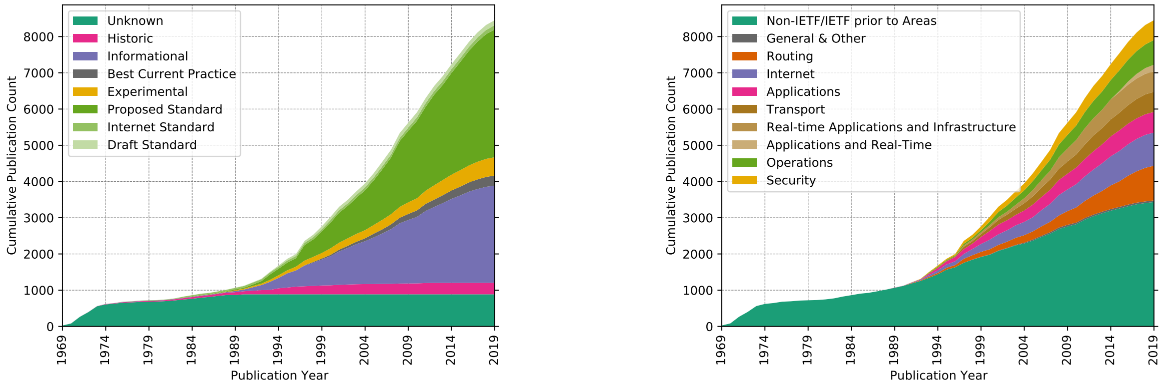
Informational RFCs have been the second largest group of publications since the early 1990s. This category includes those RFCs giving the design rationale for standards track specifications, operational guidance, deployment guidelines and considerations, as well as those defining the IETF, IRTF, and IAB processes and procedures, the role of the RFC Editor and the RFC series itself, and "April 1st" satirical publications. Smaller numbers of documents define experimental protocols and best current practices.

The highest number of RFC publications in a single year was in 2005, at the peak of the standardisation of SIP and related standards for voice-over-IP and Internet Telephony. The rate of publication has slowed in the last few years, following the completion of large work programmes relating to HTTP/2 and WebRTC.

Figure 1b shows the cumulative number of RFCs published, subdivided by IETF area, where applicable. The introduction of areas for IETF RFCs is observed around 1990, as a way of organising the increasing numbers of working groups. Areas remain relatively stable over time, with the most notable shifts being the creation of the Real-time Applications and Infrastructure area from a subset of the Transport Area, and its later merger with the Applications Area to became the Applications and Real-Time area around 2014. Another important trend that can be observed is the significant growth in the number of RFCs from the IETF Routing Area, owing to the ongoing development of standards for MPLS, service function chaining, and fat tree routing in data centres.

3.1 Citation Trends

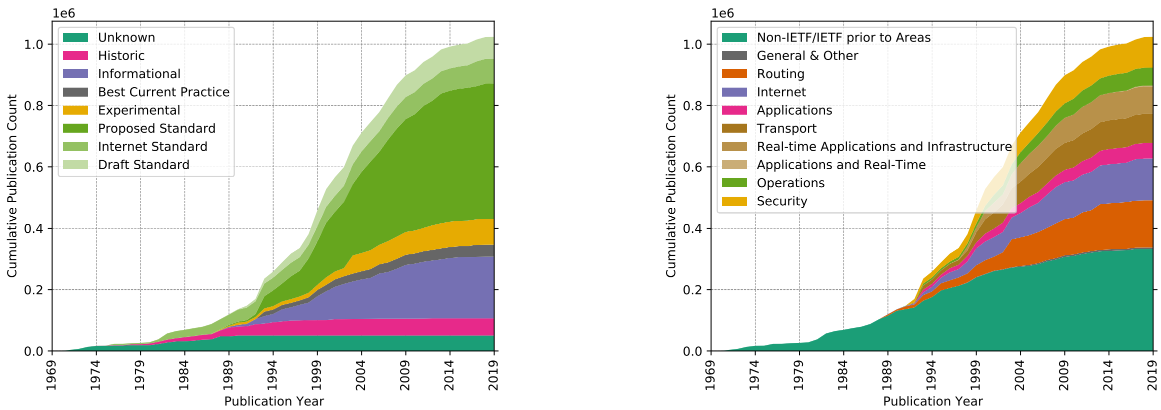
Figure 2a shows the cumulative number of citations received by each category of RFCs using citation counts from Google Scholar. We assign the number of citations of each RFC to the



(a) Cumulative RFCs by category

(b) Cumulative RFCs by IETF area

Figure 1: Cumulative number of RFCs published (1969–2019)



(a) Cumulative RFC citations by category

(b) Cumulative RFC citations by IETF area

Figure 2: Cumulative RFC citation count (1969–2019)

year in which it was published. For the period from the mid-1990s to the mid-2000s, citations to RFCs grow faster than the rate of publication of RFCs, reflecting massive growth in the networking research community as the Internet achieved globally ubiquitous deployment, and as Internet technologies began to dominate also in areas such as telephony, video streaming, and mobile devices. That growth has now slowed relative to the RFC publication rate in the last decade or so, perhaps showing that the research community has to some extent moved on from studying Internet technologies.

The standards track draws disproportionate attention in terms of citations, compared with the relative proportion of standards track RFCs, reflecting the importance of those RFCs that are standards for the operation of the Internet. In contrast, Informational RFCs receive proportionally fewer citations. This is partially due to the administrative nature of the content contained by many of these documents, which is essential to the operation of the IETF standards process but of limited interest to the broader community.

Figure 2b shows the cumulative number of citations received by RFCs, separated according to area of the IETF and non-IETF documents. Comparing to Figure 1b, it is clear that non-IETF documents are under-cited relative to their prevalence in the RFC series. This may be because older documents are less relevant to new research, but there is a potential concern that work published on the IRTF, IAB, and independent streams is insufficiently visible to the community.

The most highly cited RFCs are overwhelmingly published through the IETF stream. Within that stream RFCs produced by the Internet Area, Routing Area, Real-time Applications and Infrastructure Area (RAI), and Security Area predominate. Interestingly, RFCs from the Transport Area, that developed TCP, do not appear to be especially highly cited.

Table 1: Top Cited Internet RFCs during 1969–2019

RFC	Citations	Working Group	Area	Category
RFC3561	24017	manet	Routing	Experimental
RFC6923	14226	mpls	Routing	Proposed Standard
RFC2119	14219	No WG	No Area	Best Current Practice
RFC3550	9584	avt	RAI	Internet Standard
RFC1889	9582	avt	RAI	Proposed Standard
RFC2475	9407	diffserv	Transport	Informational
RFC2543	9039	mmusic	RAI	Proposed Standard
RFC3261	9039	sip	RAI	Proposed Standard
RFC6286	7557	idr	Routing	Proposed Standard
RFC3775	7554	mobileip	Internet	Proposed Standard
RFC3626	6251	manet	Routing	Experimental
RFC0793	6223	No WG	No Area	Internet Standard
RFC8200	5762	6man	Internet	Internet Standard
RFC1883	5748	ipngwg	Internet	Proposed Standard
RFC2460	5748	ipngwg	Internet	Draft Standard
RFC1320	5559	pem	Security	Historic
RFC1321	5559	pem	Security	Informational
RFC2205	5547	rsvp	Transport	Proposed Standard
RFC1633	5190	No WG	No Area	Informational
RFC3031	4902	mpls	Routing	Proposed Standard

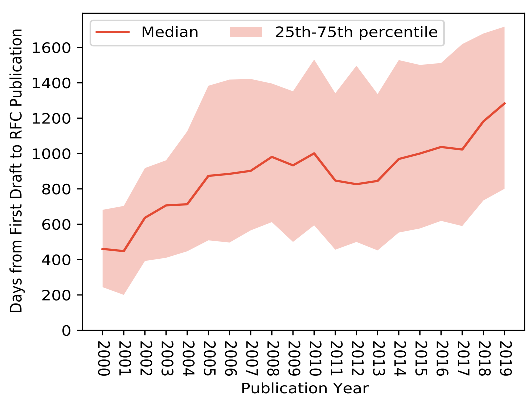


Figure 3: Time to publish for RFCs

The most cited RFCs are shown in Table 1. They include multimedia transport and signalling protocols (RTP, SIP), routing protocols (AODV, BGP, OSPF, MPLS), quality of service mechanisms (RSVP, DiffServ), TCP, and IPv6, often with multiple versions of the same protocol being widely cited. Most cited is the AODV routing protocol for mobile ad-hoc networks, an experimental RFC that has not received wide deployment. This shows a disconnect between the industry and research communities that may be worth exploring.

3.2 Complexity

As the Internet, and the community developing technical specifications for the Internet, have grown, there is a concern whether the the standards and the standards process has become more complex. In this section, we make an initial attempt to quantify the complexity of the process. To do this, we analyse the time taken for an RFC to be published, beginning from its first draft. We use data from the IETF Datatracker, looking at RFCs published between 2000–2019

Track	Page Count	References	Authors
All	0.22 (5747)	0.18 (5077)	0.18 (5747)
Standards	0.23 (3137)	0.23 (2871)	0.19 (3137)
Other	0.19 (2610)	0.14 (2206)	0.17 (2610)

Table 2: Correlation coefficients with time to publish for page counts, number of referenced documents, and number of authors (total RFCs in parenthesis).

(5747 RFCs), inclusive, where a timeline (consisting of at least one prior draft) can be constructed.

Figure 3 shows the *time to publish* for RFCs in the Datatracker dataset, defined as the number of days from submission of the first draft to the date of RFC publication. This gives a broad measure of *complexity*, encompassing the coordination costs between authors, the process of building consensus, and the editing and publication process itself.

We can see that the standardisation process is taking longer over time. The median time to publish was 460 days in 2000, and rose to 1283 days in 2019. While this encompasses many aspects of the process, we can begin to analyse the factors that may contribute to this increasing complexity.

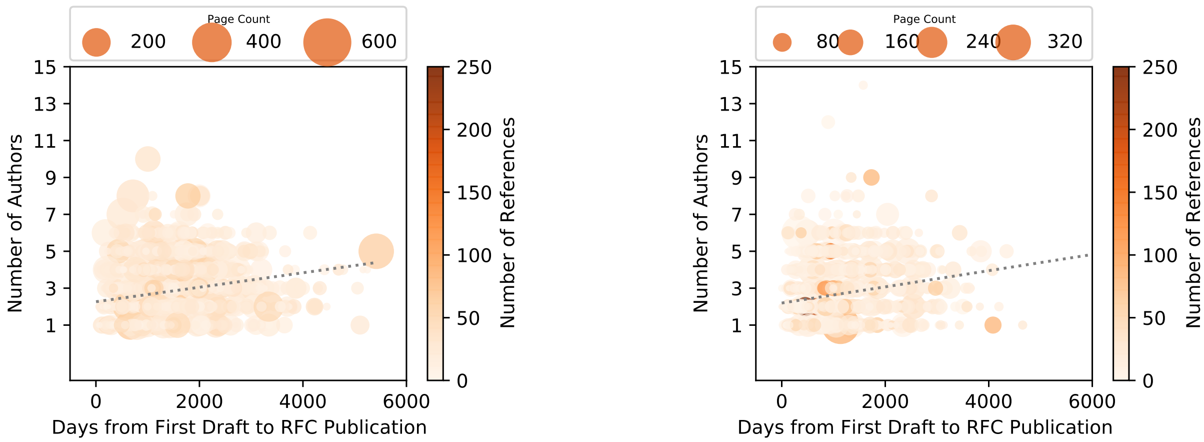
Table 2 lists the correlations between time to publish and the number of pages, referenced documents (for RFCs in the dataset that reference at least one document), and authors of each RFC.

Figure 4 further explores the relationships between each of these document attributes. We observe that while standards track RFCs typically have the same median author counts, they typically take longer to be published, reference more documents, and have more pages. In addition, we also identify weak, but statistically significant, correlations between each of the document attributes. This follows the intuition: as documents become longer, include more references, and have more authors, it takes longer for them to be published.

4 AUTHORSHIP

Given the correlation between increasing complexity and the number of authors that an RFC has, we now further explore trends in authorship.

Figure 5 shows the cumulative RFCs publications by the 5 most prolific organisations and overall by authors from academic and industrial institutions. In its early years, the RFC series was dominated by authors from academic institutions, but over time this has shifted with the greater involvement of commercial organisations starting in the early 1990s. Nowadays, the majority of RFCs have industry authors. While Cisco is still dominant, many of its contributions were made in the early 2000s, with significant work on routing standards,



(a) Standards track **(b) Other tracks**

Figure 4: Relationships between time to publish, number of authors, references, and page counts for RFCs

including MPLS. There is a marked rise in contributions from Huawei beginning around 2010.

Despite the decline in the relevance of academic organisations, a tight relationship between industry and academia remains, with top RFC publishing organisations being also very active in the main Internet Systems conferences (SIGCOMM conferences). Figure 5 shows how the top RFC publishers are also academically active regardless of whether they are industrial organisations or not.

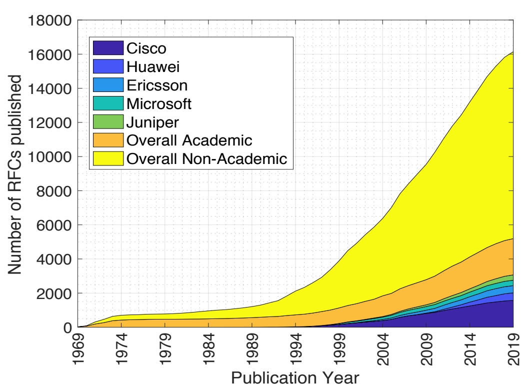


Figure 5: Top RFC publishing institutes

While corporations seek to engage in standardisation as part of their commercial agendas, RFC authorship relies upon individuals. These individuals must have pre-existing expertise, not only in their technical specialisms, but also in managing the standardisation process itself. As a result, the rise of new institutions is often caused by authors shifting from one institution to another. This is shown in Figure 7, with

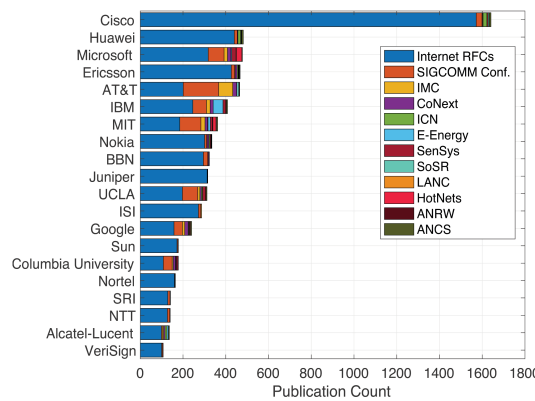


Figure 6: RFC and SIGCOMM publications for top RFC publishers

top institutions (by number of RFCs), and the flow of authors that have switched affiliations; we consider each author's first and last affiliation in the dataset. While Cisco, previously dominant, has lost a significant number of RFC authors, new incumbents, such as Huawei, have filled their ranks with experts, from Cisco (13), Ericsson (3), and others. There are many examples of authors that leave large corporations to operate private consultancy firms, reflected in the shift to "Other Non-Academic" in the figure.

Finally, we consider the geographic region where institutions are headquartered. While North America dominates the Internet standardisation process until the mid-1990s, the rise in contributions from Europe and Asia since then has been marked. This bears out the other trends in authorship

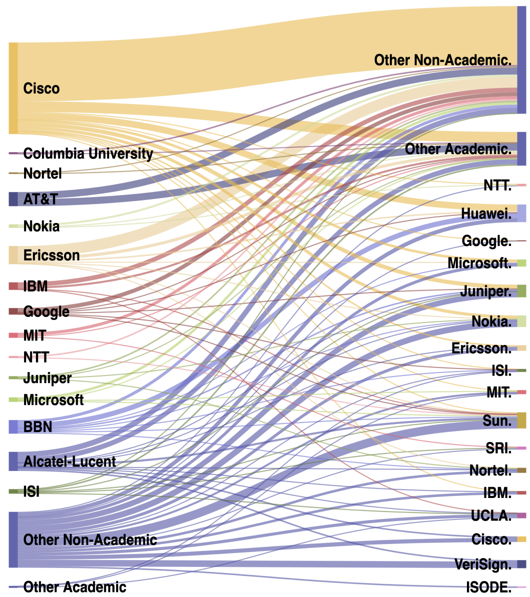


Figure 7: Change from first to last observed affiliation of RFC authors for top publishing organisations

we have identified, including, for example, the shift in dominance from Cisco to Huawei.

4.1 Collaboration

In Section 3.2, we found that the standardisation process is becoming increasingly complex. We now explore collaborations between authors as a possible contributing factor. Figure 8 shows a clear increasing trend in the median number of authors, organisations and expertise (in terms of working groups and areas) involved in RFC authorship over time. The number of authors listed for a given RFC is generally capped at 5, meaning that the numbers given in this section are lower bounds. That said, the median number of authors per RFC has tripled over time. We make two observations about the increase in the median number of authors. First, consensus from a larger number of authors, while ostensibly leading to better standards, is likely to take longer; this is supported by the preliminary results in Section 3.2.

In addition to an increase in the median number of authors, we also observe that the median number of areas, working groups, institutions, and countries represented in the authorship of each RFC also increases significantly over time. Each of these represents further sources of complexity, with the overhead of conducting collaborations between groups of different expertise, corporate or social cultures, and even timezones and languages. It remains to further explore each of these factors, and the contribution, if any, that each makes to the complexity of the standardisation process.

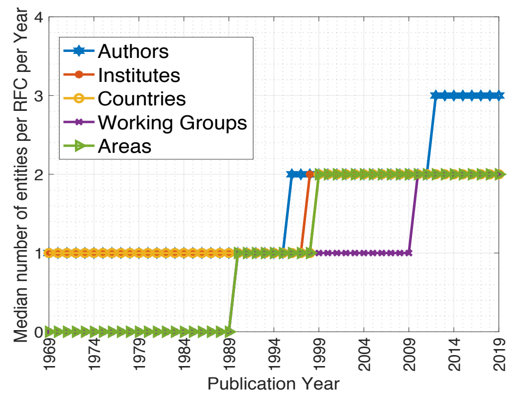


Figure 8: Median number of authors, institutes, countries, working groups, & areas in RFCs for 1969–2019

5 RELATED WORK

To the best of our knowledge, we are the first to analytically study the RFC series. That said, there have been a number of related works performing bibliometric analysis in general areas of computer science and networking [2, 4, 13, 14, 16–18]. Flittner et al. have performed quantitative analysis on selected top ACM conferences to understand publication trends [6]. The content of various research areas using techniques such as topic modeling [15] and keyword-based analysis [3] has been well studied.

Relatively limited work has focused on using bibliometric techniques to analyse the publication other areas of computer networks. Chiu et al. [2] have performed an analysis of author productivity in networking venues in 2010. Iqbal et al. studied top networking conferences and journals in [12] and analyzed 50 years of publications in SIGCOMM sponsored research venues in [11].

6 CONCLUSIONS

Internet standards have a critical socioeconomic role in underpinning the Internet, and ensuring its interoperability. In this paper, we have shown that this process is becoming increasingly diverse, with more authors, from more regions, organisations and areas of expertise. While this is beneficial to the process, we have also shown that it increases its complexity: RFCs are taking longer to publish than ever before. While we have provided initial insights into the growing complexity of the process, the importance of standardisation means that we must seek to better understand *how* decisions are made, such that the benefits of diversity can be maintained. We are particularly keen to merge the RFC dataset with other sources, such as emails and agendas to better understand the dynamics of the social process.

REFERENCES

[1] B. E. Carpenter and C. Partridge. 2010. Internet Requests for Comments (RFCs) as Scholarly Publications. *Computer Communication Review* 40, 1 (Jan. 2010), 31–33. https://doi.org/10.1145/1672308.1672315

[2] Dah Ming Chiu and Tom ZJ Fu. 2010. Publish or perish in the internet age: a study of publication statistics in computer networking research. *ACM SIGCOMM Computer Communication Review* 40, 1 (2010), 34–43.

[3] Jinho Choi, Sangyoon Yi, and Kun Chang Lee. 2011. Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. *Information & Management* 48, 8 (2011), 371–381.

[4] João M Fernandes and Miguel P Monteiro. 2017. Evolution in the number of authors of computer science publications. *Scientometrics* 110, 2 (2017), 529–539.

[5] H. Flanagan. 2019. RFC 8700: Fifty Years of RFCs. *IETF RFC-editor* (2019).

[6] Matthias Flittner, Mohamed Naoufal Mahfoudi, Damien Saucez, Matthias Wählisch, Luigi Iannone, Vaibhav Bajpai, and Alex Afanasyev. 2018. A Survey on Artifacts from CoNEXT, ICN, IMC, and SIGCOMM Conferences in 2017. *ACM SIGCOMM Computer Communication Review* 48, 1 (2018), 75–80.

[7] Russ Housley and Harald T. Alvestrand. 2009. IESG Procedures for Handling of Independent and IRTF Stream Submissions. RFC 5742. (Dec. 2009). https://doi.org/10.17487/RFC5742

[8] Russ Housley, D Crocker, and E Burger. 2011. Reducing the standards track to two maturity levels. *IETF RFC 6410* (2011).

[9] IAB and Leslie Daigle. 2007. Process for Publication of IAB RFCs. RFC 4845. (July 2007). https://doi.org/10.17487/RFC4845

[10] IAB and Leslie Daigle. 2007. The RFC Series and RFC Editor. RFC 4844. (July 2007). https://doi.org/10.17487/RFC4844

[11] Waleed Iqbal, Junaid Qadir, Saeed-Ul Hassan, Rana Tallal Javed, Adnan Noor Mian, Jon Crowcroft, and Gareth Tyson. 2019. Five decades of the ACM special interest group on data communications (SIGCOMM) a bibliometric perspective. *ACM SIGCOMM Computer Communication Review* 49, 5 (2019), 29–37.

[12] Waleed Iqbal, Junaid Qadir, Gareth Tyson, Adnan Noor Mian, Saeed-ul Hassan, and Jon Crowcroft. 2019. A bibliometric analysis of publications in computer networking research. *Scientometrics* 119, 2 (2019), 1121–1155.

[13] S Nattar. 2009. Indian journal of physics: A scientometric analysis. *International Journal of Library and Information Science* 1, 4 (2009), 043–61.

[14] Gustavo Cattelan Nobre and Elaine Tavares. 2017. Scientific literature analysis on big data and internet of things applications on circular economy: a bibliometric study. *Scientometrics* 111, 1 (2017), 463–492.

[15] Michael Paul and Roxana Girju. 2009. Topic modeling of research fields: An interdisciplinary perspective. In *Proceedings of the International Conference RANLP-2009*. 337–342.

[16] Periyaswamy Rajendran, R Jeyshankar, and B Elango. 2011. Scientometric analysis of contributions to Journal of Scientific and Industrial Research. *International Journal of Digital Library Services* 1, 2 (2011), 79–89.

[17] Alexander Serenko, Nick Bontis, and Joshua Grant. 2009. A scientometric analysis of the proceedings of the McMaster world congress on the management of intellectual capital and innovation for the 1996-2008 period. *Journal of Intellectual Capital* 10, 1 (2009), 8–21.

[18] Zhifeng Yin and Qiang Zhi. 2017. Dancing with the academic elite: a promotion or hindrance of research production? *Scientometrics* 110, 1 (2017), 17–41.