The Unsolvable Problem or the Unheard Answer? A Dataset of 24,669 Open-Source Software Conference Talks

Kimberly Truong,[1] Courtney Miller,[2] Bogdan Vasilescu,[2] Christian Kästner[2]
[1] Oregon State University, USA [2] Carnegie Mellon University, USA
truonkim@oregonstate.edu,{cmiller,vasilescu}@cmu.edu

ABSTRACT

Talks at practitioner-focused open-source software conferences are a valuable source of information for software engineering researchers. They provide a pulse of the community and are valuable source material for grey literature analysis. We curated a dataset of 24,669 talks from 87 open-source conferences between 2010 and 2021. We stored all relevant metadata from these conferences and provide scripts to collect the transcripts. We believe this data is useful for answering many kinds of questions, such as: What are the important/highly discussed topics within practitioner communities? How do practitioners interact? And how do they present themselves to the public? We demonstrate the usefulness of this data by reporting our findings from two small studies: a topic model analysis providing an overview of open-source community dynamics since 2011 and a qualitative analysis of a smaller community-oriented sample within our dataset to gain a better understanding of why contributors leave open-source software.

1 INTRODUCTION

Many researchers and practitioners have lamented the disconnect between practitioners and researchers in software engineering. Questions have been raised about whether software engineering research is still relevant and how we can close the gap between software engineering in practice and academia. For example, past research by Lo et al. [19] found there was no direct correlation between the perceived relevance of a conference paper by practitioners and its number of citations (often used in academia to determine the success of a paper). Similarly, Begel and Zimmermann [4] studied how software developers at large companies rated existing software engineering conference papers based on relevance to their work, replicated subsequently by Huijgens et al. [15]. They reported over 140 questions relevant for software engineering practitioners, among which many were not commonly explored in academia.

We take a closer look at open-source software, where the disconnect is even more of an issue. There's a direct correlation between the success of an open-source project and its community interactions [8, 10, 18]. Open-source is essential to software engineering, impacting 78% of US companies as of 2015 [12]. We need to better understand how contributors interact, what information is shared, and how communities are managed to bridge the gap between open-source practitioners and researchers and sustain open-source.

We approach this issue by curating a dataset of 24,669 open-source software conference talks to understand practitioner perspectives. Conference talks are a source of grey literature [13], which is a large and mostly untapped data source providing first-hand accounts from the practitioners themselves. These accounts reveal what open-source software practitioners talk about and what they find important, rather than just what researchers surmise are key topics to study. These conference talks are available online so the data collection is less costly than surveys or interviews while still providing value. With this dataset, we gain insight into what open-source software communities care about, what they want to share, and how they present themselves to the public. Furthermore, grey literature documents the events and intent of the speaker at the time of recording, capturing how open-source software has evolved over time (one decade, Jan. 2011 - June 2021, in our dataset). Many talks have been recorded and uploaded to YouTube (with over 500 open-source software talks in 2011 and 11,000 talks in 2020). We included all conferences related to open-source software by searching through the most popular results on Google and online databases (e.g., calendars, Wikipedia pages). These conference talks range from how open-source communities function and issues that need to be addressed to technical project updates and new ideas that should be pursued. Our dataset is broad and diverse including speakers from 87 conferences of various sizes, locations (e.g., United States, United Kingdom, Australia), disciplines, roles (e.g., maintainer, contributor, user), and times creating a representative dataset of open-source software communities as a whole.

Past research regarding open-source practitioners mainly focused on what makes a project successful, how to encourage new contributors to join, and why contributors leave open-source software. These typically discuss specific problems or topics (e.g., mailing lists [14], GitHub issue threads [2, 21]) and any interactions with practitioners are only feasible with small to moderate sample sizes due to the cost of data collection (e.g., interviews [16], surveys [7]). These studies found that practitioners most commonly talk about implementation problems and project comprehension. Our dataset expands on these studies by including a diverse sample of practitioner perspectives demonstrating how they communicate and what they want to share.

In addition to our dataset, we provide a tool to extract YouTube video data, including video transcripts which can be used for further grey literature analysis. Videos present insight into a person's experiences where they control the narrative [6]. We can expand this tool beyond just open-source to other sub-fields of software engineering or even beyond conferences.

We are enthusiastic about future research in this area. We want to use this dataset to answer questions such as: What do practitioners

MSR '22, May 23–24, 2022, Pittsburgh, P A, USA

ACM ISBN 978-1-4503-9303-4/22/05.
https://doi.org/10.1145/3524842.3528488

want to talk about? How do ideas or technologies spread through open-source communities? How does speaking at a conference impact the practitioner and the project? And how has open-source evolved and where is it heading in the future? We begin answering these questions with two applications we briefly demonstrate in this paper: a topic model analysis and a qualitative analysis of why contributors leave open-source software.

2 METHODOLOGY

We identify talks that have been recorded at open-source software conferences and uploaded to YouTube. Many open-source software conferences have a long tradition of recording and releasing talks (more than academic conferences that primarily have papers in proceedings as the primary material). To identify relevant talks, we pinpoint YouTube channels that contain conference talks. The process of identifying relevant conferences and corresponding channels was mostly performed manually, following a deliberate process and rubric. A key step in this process is to assure data quality by judging relevance of conferences for our corpus.

Once we identified a corpus of channels containing relevant talks (from 87 conferences), we downloaded metadata and transcripts.

Conference list. We create a list of conferences with topics relating to open-source. To identify our corpus of conferences, first, we considered the top 30 search choices on Google with the keywords 'open source conference' and 'open source conference call for proposals.' Second, we looked through each link and noted any conference names mentioned on the link or within a couple of clicks from the link. We added the conference name to our list. Useful links included calendars of free-and-open-source (FOSS) events and the *open-source software conferences Wikipedia page.*[1] These calendars contained a list of open-source software conferences that happened in recent years and are planned for the next year. We looked back two years to May 2019 on each of these calendars and collected all conference names from the lists.

Constraints and Metadata. There are many talks related to open-source that can be found on YouTube, coming from many conferences. Conferences differ substantially in size and in how systematically they upload talks (e.g., all talks, just keynotes, or just some authors who decide to record and upload their talks). To assure our dataset is as representative of open-source communities as possible, we filtered the conferences considered with the following constraints:

- There are at least two documented editions (with 3 or more recorded talks) accessible through YouTube. This ensures there are relevant data about the conference to collect.
- The conference is a notable size and has some impact on the community (where notable size means it has at least 50 attendees, or 10 speakers/talks).

These conditions assure the conferences in our dataset have some impact on the community and have enough accessible data to be used for further analyses. We further filter for data analysis by excluding conferences that aren't in English to increase the chances of an existing (or auto-generated) transcript on YouTube.

We check if our conferences meet these constraints by manually collecting metadata about each one and its documented editions. Our metadata includes: focus/theme, size (number of talks, speakers, and attendees), affiliated conferences/organizations, sponsorship information, main website (or most recent website), and Wikipedia page. We found this information by searching Google for the conference name. We looked at the first four websites to find relevant information and stopped once we found all the information we needed. Then we searched for the conference name + 'Wikipedia' and for each edition in descending order to get more edition-specific information. We considered any editions after 2011 but mainly focused on the editions we could access through YouTube. We share this metadata with our dataset. After filtering, we arrived at a list of 87 conferences that match our filtering criteria.

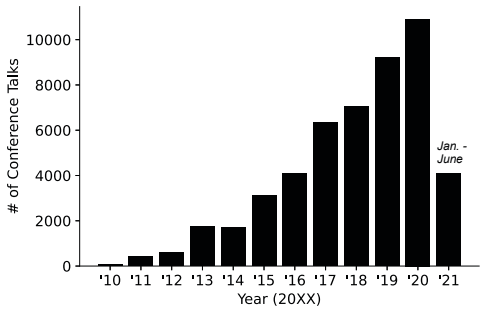

Figure 1: Distribution of conference talks in dataset over time (2010 - June 2021)

Data Compilation. After identifying relevant conferences and their YouTube channels, we collected all talks from those channels. Then, we parsed through each conference's YouTube channel (or some playlist containing the conference talks) with our scripts using the Google Data API and the PyTube library. These scripts downloaded information regarding all the conference talks for each conference in our list by parsing through all the playlists on the YouTube channel. Often each playlist is a different conference edition. From there, we created a directory for each conference and then sub-directories for each conference edition (this was given since most channels had their editions separated into playlists). Within these sub-directories, each conference talk/video had its own text file containing the:

- Name of the video
- Publication date
- Playlist (often the conference edition)
- Description
- Transcript (given by the scripts)
- YouTube URL

We created scripts to combine this dataset into a large csv file for further processing and applications.

3 DATASET DESCRIPTION

The dataset is available at DOI 10.5281/zenodo.6395342 [22]. Due to the YouTube license, we only share channel IDs, metadata, and scripts, but not actual transcripts of the talk. It is easy to download the transcripts, descriptions, and even the full videos for further analysis using the scripts we release.

[1] https://en.wikipedia.org/wiki/List_of_free-software_events

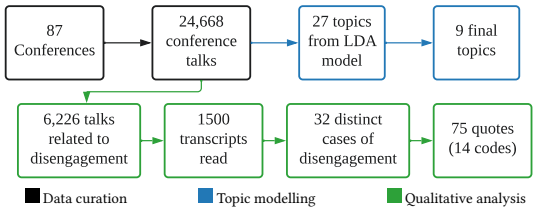

Figure 2: Data collection and processing

Our dataset consists of 24,669 conference talks from 87 open-source software conferences ranging from 2010 to 2021. For most applications, we would like to consider the range 2011 to 2020 since 2010 only has 84 talks and 2021 is an incomplete sample covering only January to June (the time when we collected the data). As illustrated in Figure 1, as time passes, there is an increase in conference talk recordings, which coincides with the increasing use of digital media and ease of recording and publishing recordings.

Our dataset contains conferences of various sizes and disciplines. We have major conferences like *PyCon*, *JSConf*, and the *Linux Cloud Conference*, as well as lesser-known conferences like *BazelCon* and *GitOpsCon*. This provides a lot of variety in our dataset and many interesting possible applications which can be done to compare and contrast these conferences. It contains talks covering many different topics, technical and societal. These topics include operating systems, programming language features, community interactions, company management, and governmental interference. This dataset represents how practitioners view open-source software as a whole. The talks provide insight on practitioner workflow, application uses, project setbacks/issues, interests, project sponsors, and more.

4 APPLICATIONS

This dataset can be used for a variety of applications such as: analyzing the popularity and diffusion of tools and practices in open-source, identifying who/what has the greatest influence on certain communities, understanding common challenges discussed by practitioners, filtering talks by dates to identify how certain events impacted open-source communities, analyzing trends in open-source conferences over time, by focus, and by size, and analyzing the similarities and differences between conferences.

As a demonstration of the value of this dataset, we briefly report findings from two applications to gain a better understanding of open-source community dynamics including interactions, goals, and expectations. We start by performing a topic model analysis and follow with thematically coding the top talks related to contributors leaving open-source. Our topic model analysis provides an overview of what people talk about and how this dynamic has changed over time; our second application uses a qualitative approach to analyze a sample from our corpus matching a list of keywords.

4.1 Topic modeling

We use our topic model analysis to obtain conference themes (beyond just major projects or conference names) and a list of talks in each topic. Understanding what topics open-source encompasses lets us identify niche sub-topics in each one and validate our model. This analysis is inspired by a topic model analysis of software-engineering conference paper abstracts done by Mathew et al. [20]. We first pre-process our corpus to filter out irrelevant text (e.g., filler words, uncommon names). This lets our data create an overview including both more niche topics (e.g., *Workflow*), as well as common, but not overused topics (e.g., *Python Applications*, *Privacy*). We pre-process our corpus by running a term frequency–inverse document frequency (tf-idf), where we consider words with a frequency under 0.002 to be irrelevant and remove them during the topic model analysis. Then, we input our data to a Latent Dirichlet Allocation (LDA) Model with 27 topics. We tested a range of 7 to 40 topics and found that 27 topics had the best inter- and intra-similarity rates.

Interpretation. Each topic is represented by a distribution over words [5, 17]. We interpret a list of the 30 most common words and conference talks in each topic. The keywords in these talks indicate their top topic. We use these factors to compare how conferences in a topic are similar and how the 30 most common words differ from words in other topics. We name each topic based on these differences. Then, we manually consolidate these 27 topics based on word and title similarity from the talks in each topic [20]. An example of similar titles is *Women Representation and Mentoring* and *Diversity, Leadership, and Community Interactions*. We consolidated these into *Community Interactions*. After consolidating the topics, we found 9 distinct open-source software topics. We list the topics below with notable keywords and the common conferences found in each one.

* **Databases**; query, render, sql; *GraphQL Summit, Berlin Buzzwords*
* **Game development**; player, consequence, animation; *EuroPython , JSConf*
* **Workflow**; community, pull (request), integration; *RubyConf, LinuxFest*
* **Lessons and Mentoring**; teach, feedback, sponsor; *RustFest, CppCon*
* **Security**; monitoring, token, validation; *Open Source Summit, KVM Forum*
* **UX/UI Design**; web, content, users; *CppCon, JSConf*
* **Privacy**; government, test, vulnerability; *FOSS Backstage, State of the Map*
* **Community Interactions**; diversity, community, users; *All Things Open, JupyterCon*
* **Python Applications**; application, cloud, python; *ApacheCon, PyCon*

We find most of these talks are technical, discussing programming languages, tool use, and how the project will evolve in the future; A couple of topics were related to open-source regulation and management. Most notably, we see many talks discuss social issues, preferences, and relationships (e.g., *Community Interactions*, *Lessons and Mentoring*). The topics have not changed in the past decade and the distribution of talks in each topic has remained fairly stagnant (see Figure 3), with the exception of *Community Interactions*. Talks relating to *Community Interactions* have increased every year (with a large increase between 2017 and 2018). These talks include encouraging people to contribute to open-source, discussing issues practitioners noticed in open-source, and sharing how practitioners would like open-source to change in the future.

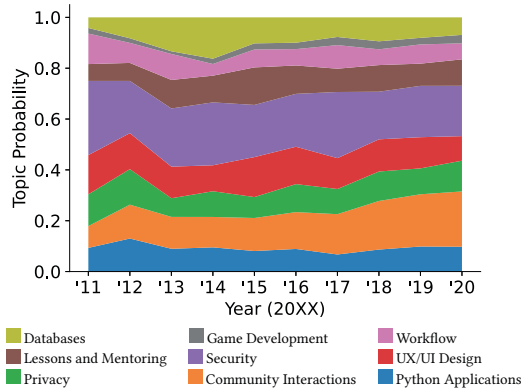

Figure 3: Probability of OSS topics over time (2011-2020)

This trend raises new questions. Why was there a sudden increase in 2017? How have more talks regarding *Community Interactions* affected open-source communities as a whole? And how do conferences that discuss social aspects differ (if at all) from those that do not? We find in our second application that conference speakers share their reasons for leaving open-source and often give advice for how they believe open-source software could be more sustainable. Thus, increasing discussions of *Community Interactions* is a positive trend for open-source software.

4.2 Analyzing a Disengagement Sample

We take a more in-depth look at a major issue in open-source sustainability — disengagement. We define this as when a contributor either pauses their project for over 3 months, leaves their project, or quits all open-source projects. Most open-source projects are dependent on a small group of core contributors. A study by Avelino et al. [3] on the 133 most popular GitHub applications found that 86% of projects are likely to fail if one or two of its core contributors leave. We want to understand how contributors and users interact through a brief overview of how the community present themselves and discuss their experiences with disengagement. We aim to gain a better understanding of how to improve open-source communities and prevent contributors from leaving by listening to the challenges cited by practitioners when discussing disengagement.

Data analysis. We generate a sample from our dataset by filtering with keywords such as 'leave', 'abandon', 'hostility', and more. We selected these keywords based on past studies [7, 11, 21] and known talks [1] on open-source disengagement.

We analyze this sample by sorting the talks in descending order by the number of keyword occurrences. We manually skim through the top 1500 transcripts and thematically code 34 relevant ones (with two duplicate speakers) by: 1. Recording quotes related to disengagement 2. Having two researchers read over the quotes to identify codes 3. Combining this data with another researcher investigating reasons for disengagement cited in blog posts. 4. Generating a code book based on the new data 5. Having two researchers read over each quote again and assign final codes.

Results. We found three major categories of reasons cited for disengagement (each with 4-5 codes): volunteering-related (50%), cultural (32%), and external (18%). The most common reasons among conference speakers were lack of support (emotionally and financially, with 10 cases citing lack of compensation) and community hostility. These findings are documented at http://disengagement-diaries.github.io. The website also stores each disengaged contributor's contact information, their reasons for disengagement, and their recommended interventions to prevent future contributor disengagement. The most commonly recommended interventions were to encourage and maintain a work-life balance and to promote more inclusive communities. Our results support previous studies regarding the importance of community engagement in open-source [8, 9] and more importantly, the responsibility of community members to support their peers. It's especially important to support other contributors to prevent disengagement from open-source software and to continue to sustain open-source projects.

5 LIMITATIONS

Users of this dataset should be aware of the YouTube license, the possible inconsistency between the date the talk was given and the publication date, and the exclusion criteria placed on all conferences in the dataset: First, we do not share the video transcripts directly in our dataset, but instead provide all relevant metadata (see Section 2) and the channel ID which can be input into our scripts to collect all video information (including transcripts) for that conference. This process is straightforward and only requires calling one function. Second, in the data collection process, we also collect the publication date. This date is not completely accurate of when the conference took place. For our topic model analysis, we use this interchangeably with the time of the conference and categorize all the conferences by year for our timeline. A manual inspection found that conference talks are usually uploaded to YouTube within a year of the actual event, so we believe the reported timeline to represent actual time is reasonably accurate. Finally, the use of constraints assured the quality of our data but required manual collection of the metadata. This could have resulted in some conferences being excluded if such metadata could not be found with our systematic approach. Thus, users of our dataset should be careful when generalizing beyond the exclusion criteria provided in Section 2.

6 CONCLUSIONS

We curated a dataset containing 24,669 open-source software conference talks with metadata from 87 conferences and provided a tool to collect YouTube video transcripts. These talks are all first-hand accounts from practitioners and are representative of open-source software in practice showing the evolution of practitioner interests, workflows, and community dynamics over time. This dataset can be used to identify how conferences affect open-source projects, how talks differ by discipline/topic within open source, and how practitioners recommend we promote open-source sustainability. We hope our dataset helps future work bridge the disconnect between practitioners and researchers and improve open-source software based on community recommendations to increase open-source sustainability.

Acknowledgements. This work was supported by a grant from the Sloan Foundation.

REFERENCES

[1] PyCon 2019. 2019. *Russell Keith-Magee - Keynote - PyCon 2019.* YouTube. https://youtu.be/ftP5BQh1-YM?t=3000

[2] Deeksha Arya, Wenting Wang, Jin LC Guo, and Jinghui Cheng. 2019. Analysis and detection of information types of open source software issue discussions. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 454–464.

[3] Guilherme Avelino, Marco Tulio Valente, and Andre Hora. 2015. What is the Truck Factor of popular GitHub applications? A first assessment. *PeerJ Preprints* 3 (2015). https://doi.org/10.7287/peerj.preprints.1233v3

[4] Andrew Begel and Thomas Zimmermann. 2014. Analyze This! 145 Questions for Data Scientists in Software Engineering. In *Proceedings of the 36th International Conference on Software Engineering* (Hyderabad, India) *(ICSE 2014)*. Association for Computing Machinery, New York, NY, USA, 12–23. https://doi.org/10.1145/2568225.2568233

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. In *Journal of Machine Learning Research*. 993–1022.

[6] Souti Chattopadhyay, Denae Ford, and Thomas Zimmermann. 2021. Developers Who Vlog: Dismantling Stereotypes through Community and Identity. *CoRR* abs/2109.06302 (2021). https://arxiv.org/abs/2109.06302

[7] Jailton Coelho and Marco Tulio Valente. 2017. Why Modern Open Source Projects Fail. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2017)*. 186–196. https://doi.org/10.1145/3106237.3106246

[8] Kevin Crowston and Ivan Shamshurin. 2017. Core-periphery communication and the success of free/libre open source software projects. *Journal of Internet Services and Applications* 8, 10 (2017). https://doi.org/10.1186/s13174-017-0061-4

[9] Sherae Daniel, Ritu Agarwal, and Katherine J Stewart. 2013. The effects of diversity in global, distributed collectives: A study of open source project success. *Information Systems Research* 24, 2 (2013), 312–333.

[10] Nicolas Ducheneaut. 2005. Socialization in an open source software community: A socio-technical analysis. *Computer Supported Cooperative Work (CSCW)* 14, 4 (2005), 323–368.

[11] Nadia Eghbal. 2020. *Working in public: the making and maintenance of open source software.* Stripe Press.

[12] Ellak. [n.d.]. The Ninth Annual Future of Open Source Survey. https://gfoss.eu/the-ninth-annual-future-of-open-source-survey/. Accessed: 2021-12-21.

[13] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. In *Information and Software Technology*. 101–121. https://doi.org/10.1016/j.infsof.2018.09.006

[14] Anja Guzzi, Alberto Bacchelli, Michele Lanza, Martin Pinzger, and Arie van Deursen. 2013. Communication in open source software development mailing lists. In *2013 10th Working Conference on Mining Software Repositories (MSR)*. 277–286. https://doi.org/10.1109/MSR.2013.6624039

[15] Hennie Huijgens, Ayushi Rastogi, Ernst Mulders, Georgios Gousios, and Arie van Deursen. 2020. *Questions for Data Scientists in Software Engineering: A Replication.* Association for Computing Machinery, New York, NY, USA, 568–579. https://doi.org/10.1145/3368089.3409717

[16] Giuseppe Iaffaldano, Igor Steinmacher, Fabio Calefato, Marco Gerosa, and Filippo Lanubile. 2019. Why do developers take breaks from contributing to OSS projects? A preliminary analysis. *arXiv preprint arXiv:1903.09528* (2019).

[17] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, and Yanchao Li. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 88, 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

[18] Rajdeep Kaur, Kuljit Chahal Kaur, and Munish Saini. 2020. Understanding community participation and engagement in open source software Projects: A systematic mapping study. In *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2020.10.020

[19] David Lo, Nachiappan Nagappan, and Thomas Zimmermann. 2015. How Practitioners Perceive the Relevance of Software Engineering Research. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering* (Bergamo, Italy) *(ESEC/FSE 2015)*. Association for Computing Machinery, New York, NY, USA, 415–425. https://doi.org/10.1145/2786805.2786809

[20] George Mathew, Amritanshu Agrawal, and Tim Menzies. 2017. Trends in Topics at SE Conferences (1993-2013). In *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*. ICSE, 397–398. https://doi.org/10.1109/ICSE-C.2017.52

[21] Courtney Miller, David Gray Widder, Christian Kästner, and Bogdan Vasilescu. 2019. Why do People Give Up FLOSSing? A Study of Contributor Disengagement in Open Source. In *IFIP International Conference on Open Source Systems*. IFIP, 116–129.

[22] Kimberly Truong, Courtney Miller, Bogdan Vasilescu, and Christian Kästner. 2022. OSS Conference Talks Dataset. https://doi.org/10.5281/zenodo.6395342