Let's Talk Open-Source — An Analysis of Conference Talks and Community Dynamics

Kimberly Truong
Oregon State University, USA

1 INTRODUCTION

Open-source software has integrated itself into our daily lives, impacting 78% of US companies in 2015 [11]. Past studies of open-source community dynamics have found motivations behind contributions [3, 14, 18, 19] and the significance of community engagement [8, 17], but there are still many aspects not well understood. There's a direct correlation between the success of an open-source project and the social interactions within its community [7, 9, 17]. Most projects depend on a small group. A study by Avelino et al. [4] on the 133 most popular GitHub projects found that 86% will fail if one or two of its core contributors leave. To sustain open-source, we need to better understand how contributors interact, what information is shared, and what concerns practitioners have. We study common topics, how these have changed over time (2011 - 2021), and what social issues have appeared within open-source communities. Our research is guided by the following questions: (1) How is open-source changing/evolving? (2) What changes do practitioners believe are necessary for open-source to be sustainable?

Previous studies regarding open-source software communities involve analyses of conference papers [17, 20, 21], mailing lists [13], GitHub issue threads [2, 22], and surveys and interviews with contributors [4, 6, 15]. These studies were limited to a smaller sample or just one case study. They found that people most commonly talk about implementation problems and project comprehension. All these studies focus on aspects that serve a specific purpose (e.g., new applications, issues) and are limited to communication within the community, or what's shared with the public.

We expand on previous research by performing a grey literature analysis [12] on open-source software conference talks. Grey literature provides publicly-available first-hand accounts of events and captures what and how communities functioned at the time. These talks haven't been modified. They provide valuable insight into the mindset of the practitioner at the time of recording. Many talks have been recorded and uploaded to YouTube since 2011 (with over 500 open-source software talks in 2011 and 11,000 talks in 2020). We curated a dataset containing 24,669 talks from 87 conferences. We included all conferences related to open-source software by searching through the most popular results on Google and online databases. These conference talks range from how communities function and issues that need to be addressed to technical project updates and new ideas that should be pursued. The variety in talks and conferences creates a representative dataset of open-source software communities as a whole.

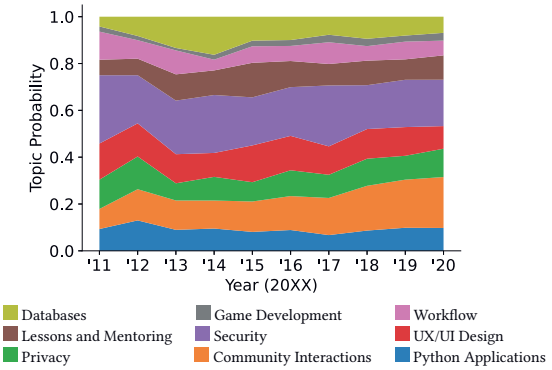

Figure 1: Probability of OSS topics over time (2011-2020)

We use this dataset to find nine topics in open-source software shown in Figure 1, with the most significant finding being the increase of talks regarding community interactions. We then qualitatively analyze talks specifically related to contributors leaving open-source communities. These speakers often share advice regarding how to improve their project or community and their hopes for what direction open-source should take in the future. Our results provide a dataset that can be further analyzed to understand different aspects in open-source communities, how open-source is changing, and how to prevent current contributors from leaving.

2 METHODOLOGY

We collect conference talks that are representative of open-source software and include the necessary information (conference title, date, and video transcript) for further analysis. We create a systematic approach to curate our dataset of 24,669 conference talks to ensure our work can be replicated and includes all relevant talks despite the variety in conference sizes, availability on YouTube, and differing conference types (e.g., purely technical, social focus, project-specific). This dataset is saved at: http://github.com/KimberlyTruong/Open-Source-Conferences.

We use our dataset to identify ways to improve open-source communities and prevent contributors from leaving by understanding community interactions and listening to the challenges cited by practitioners. 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 provides an overview of what practitioners talk about and how this dynamic has changed over time, while our second application uses a qualitative approach to analyze a sample from our corpus matching a list of keywords.

ICSE '22 Companion, May 21–29, 2022, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9223-5/22/05…$15.00
https://doi.org/10.1145/3510454.3522683

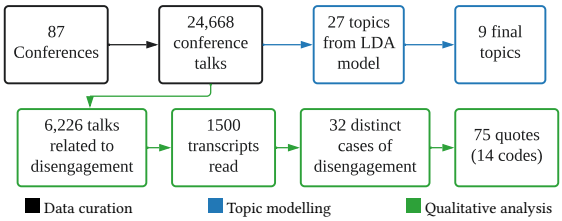

Figure 2: Data collection and processing flowchart

Curating the Dataset. We begin by manually creating a list of conferences and identifying their YouTube channels or playlists. We then collect metadata for those conferences and determine if they meet our constraints. Finally, we download all the recorded talks on YouTube. We search for conferences by considering the top 30 links on Google with the terms "open-source conference" and "open-source conference call for proposals." All conference names mentioned are noted. Useful links include calendars of free-and-open source events and the *open-source software conferences Wikipedia page.* [1] These calendars contain a list of open-source conferences from recent years. Conferences differ 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 own talks). To assure the quality of our dataset, we filter the conferences with the following constraints: The talks are in English to increase the chances of an existing (or auto-generated) transcript; There are at least two documented editions we could access through YouTube to collect more relevant data; The conference was a notable size and has had some impact on the community (where notable size means it has at least 50 attendees, or ten speakers/talks).

We verify constraints by collecting metadata about each conference and its documented editions (after 2011). After filtering, we arrive at a list of 87 conferences. We collect all talks from those channels and parse through each conference's YouTube channel using the Google Data API and PyTube library. This script downloads information regarding all the conference talks for each conference by parsing through all playlists on the YouTube channel. Often each playlist is a different conference edition. We create a directory for each conference with sub-directories for each conference edition.

Topic Model. We perform a topic model analysis on our dataset to obtain conference themes beyond just major projects or conference names and a list of talks in each topic. We first pre-process our corpus to filter out irrelevant text (e.g., filler words, uncommon names). Our data provide an overview with both niche topics (e.g., Workflow) and common, but not overused topics (e.g., Python Applications, Privacy). We pre-process 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. 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. Each topic is represented by a distribution over words [5, 16]. We use these words to name each topic and manually consolidate these based on word and content title similarity.

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

Analyzing a Disengagement Sample. We then look at a major issue in open-source – disengagement. We define this as when a contributor pauses contributions for over three months or leaves their project. Practitioners share their experiences, challenges, and recommended interventions at conferences. This highlights the changes open-source communities need to be successful.

We generate a sample of these stories by filtering from our dataset with keywords such as leave, abandon, and hostility. We select these keywords based on past studies [6, 10, 22] and known talks [1] on open-source disengagement. We sort the talks in descending order of keyword occurrences. We manually skimmed through the top 1500 transcripts and thematically coded 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 or themes 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.

3 RESULTS

Our topic model returned nine open-source software topics shown in Figure 1. Most topics involve open-source in practice and have not shifted. Our most significant finding is the growth of community interactions (with more notable growth in 2017). This is a positive trend in open-source as these talks provide support for community members and address issues that affect community engagement.

Additionally, we found three major categories 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. Each disengaged contributor's contact information, reasons for disengagement, and recommended interventions are documented at http://disengagement-diaries.github.io. We found the most common interventions were to encourage and maintain a work-life balance and to promote inclusive communities. These recommendations support previous studies on the importance of community engagement in open-source [7, 8] and the responsibility of community members to support their peers, especially contributors, to prevent disengagement and continue sustaining open-source.

4 CONTRIBUTIONS

We curated a dataset of 24,669 conference talks showcasing open-source practices and community dynamics and contributed to a database documenting reasons practitioners left open-source and their recommended interventions. We observed that conference topics have remained relatively stagnant from 2011 to 2020, except for community interactions increasing in 2017. This is a positive trend supporting what frustrated contributors cited as their recommended intervention to keep open-source sustainable: fostering supportive communities. We recommend implementing practitioner recommendations and further analyzing the dataset for trends in open-source software with the aim to improve open-source community dynamics and 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, Eleni Constantinou, Marco Tulio Valente, and Alexander Serebrenik. 2019. On the abandonment and survival of open source projects: An empirical investigation. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 1–12.

[4] 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

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

[6] 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

[7] 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

[8] 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.

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

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

[11] 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.

[12] 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

[13] 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

[14] Yu Huang, Denae Ford, and Thomas Zimmermann. 2021. Leaving My Fingerprints: Motivations and Challenges of Contributing to OSS for Social Good. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1020–1032.

[15] 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).

[16] 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

[17] 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

[18] Rajdeep Kaur and Kuljit Kaur Chahal. 2017. Analysis of Community Dynamics in Open Source Software Projects. In *International Journal of Advance Research in Science and Engineering*, Vol. 6. 1821–1830.

[19] Amanda Lee, Jeffrey C. Carver, and Amiangshu Bosu. 2017. Understanding the Impressions, Motivations, and Barriers of One Time Code Contributors to FLOSS Projects: A Survey. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*. 187–197. https://doi.org/10.1109/ICSE.2017.25

[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] Kelvin McClean, Des Greer, and Anna Jurek-Loughrey. 2021. Social network analysis of open source software: A review and categorisation. *Information and Software Technology* 130. https://doi.org/10.1016/j.infsof.2020.106442

[22] 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.