

# The Unsolvable Problem or the Unheard Answer?

## A Dataset of 24,669 Open-Source Software Conference Talks

Kimberly Truong,<sup>1</sup> Courtney Miller,<sup>2</sup> Bogdan Vasilescu,<sup>2</sup> Christian Kästner <sup>2</sup>  
<sup>1</sup> Oregon State University, USA    <sup>2</sup> Carnegie Mellon University, USA



### Motivations

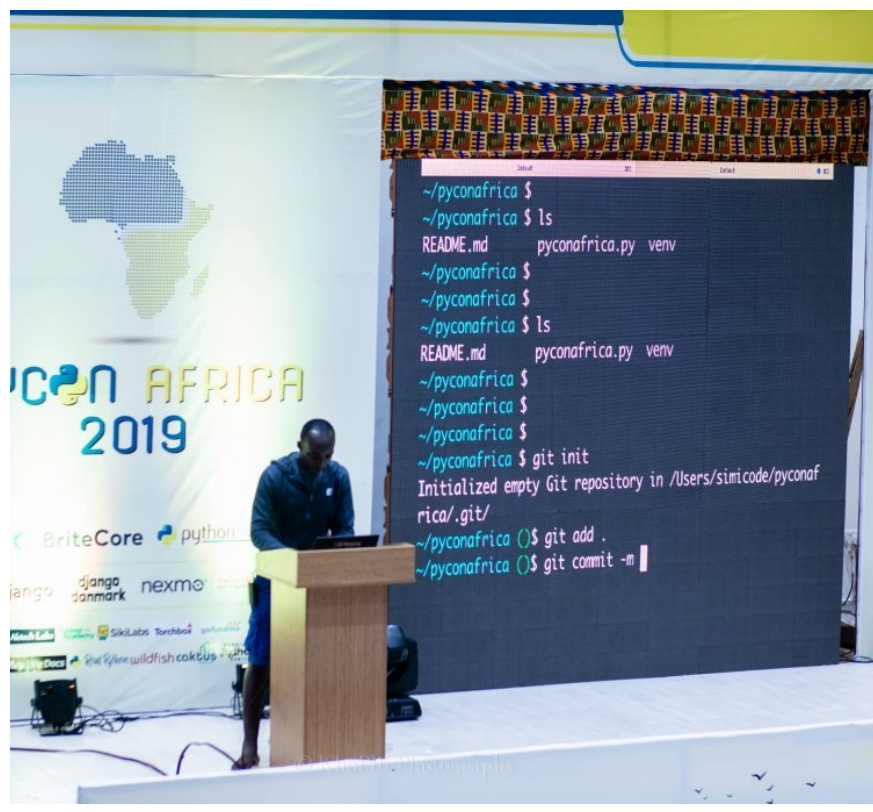
There’s a disconnect between open-source researchers and practitioners.

- There is no correlation between the perceived relevance of a software engineering conference paper and its number of citations.
- Questions relevant to practitioners were rarely found in research papers.

Community interactions in a project are an indication of the project’s success. We want to better understand practitioners through what they share.

### Source of Grey literature

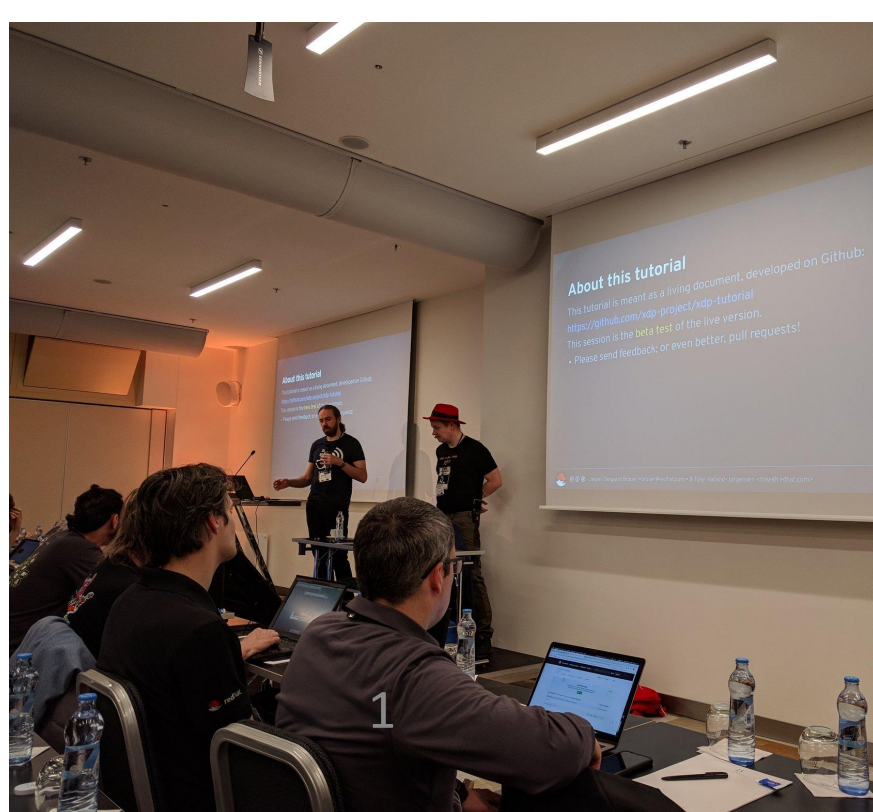
- Large and diverse untapped source(24,669 talks) publicly available online
  - Disciplines – different tools, specialties, fields
  - Size – small (50 attendees) to large (~ 4,000 attendees)
  - Experiences/roles – contributor, core maintainer, user, CEO
  - Regions – North America, Europe, Australia, parts of Africa
  - Times – 10 years (2010 - 2021)
- Provides first-hand accounts from practitioners



PyCon Africa 2019,  
<https://www.ict4d.at/tag/pycon/>



Open Source Summit 2019,  
[https://live.staticflickr.com/65535/51647514916\\_acc4a7e90b\\_c.jpg](https://live.staticflickr.com/65535/51647514916_acc4a7e90b_c.jpg)



NetDev 2019,  
<https://bootlin.com/blog/tag/netdev/>

### Dataset

#### Information

- Name of video
- Publication date
- Playlist (conference edition)
- Description
- Transcript\*
- YouTube URL

#### Metadata

- Focus/theme
- Size
  - # of talks
  - # of speakers
  - # of attendees
- Affiliated conferences/organizations,
- Sponsorship information
- Main (or most recent) website

#### Tool

- Scripts to extract YouTube video data

```
1 Title: Keynote - PyCon 2019
2 Publication date: 2019-05-05
3 Playlist: PyCon 2019 - Keynotes
4 Description:
5 "Speaker:
6
7 Keynote
8
9 Slides can be found at: https://speakerdeck.com/pycon
10 Captions:
11 00:00:41,270 --> 00:00:51,399
12 good morning
13
14 00:00:42,940 --> 00:00:53,800
15 [Applause]
16
17 00:00:51,399 --> 00:01:00,820
18 I couldn't be more excited to welcome
19
20 00:00:53,800 --> 00:01:03,699
21 you to the 2019 Pike on pike on 2019
22
23 00:01:00,820 --> 00:01:05,770
24 here in Cleveland Ohio on behalf of the
25
26 00:01:03,699 --> 00:01:08,380
27 pike on 2019 staff I want to start by
28
29 00:01:05,770 --> 00:01:09,610
30 saying thank you thank you to all the
31
32 00:01:08,380 --> 00:01:11,770
33 volunteers that make this conference
34
35 00:01:09,610 --> 00:01:13,509
36 possible to the Python Software
37
38 00:01:11,770 --> 00:01:15,820
39 Foundation for taking on the fiscal
```

\* Not shared with the Dataset due to the YouTube License

### Methodology

1. Establish a **conference list**

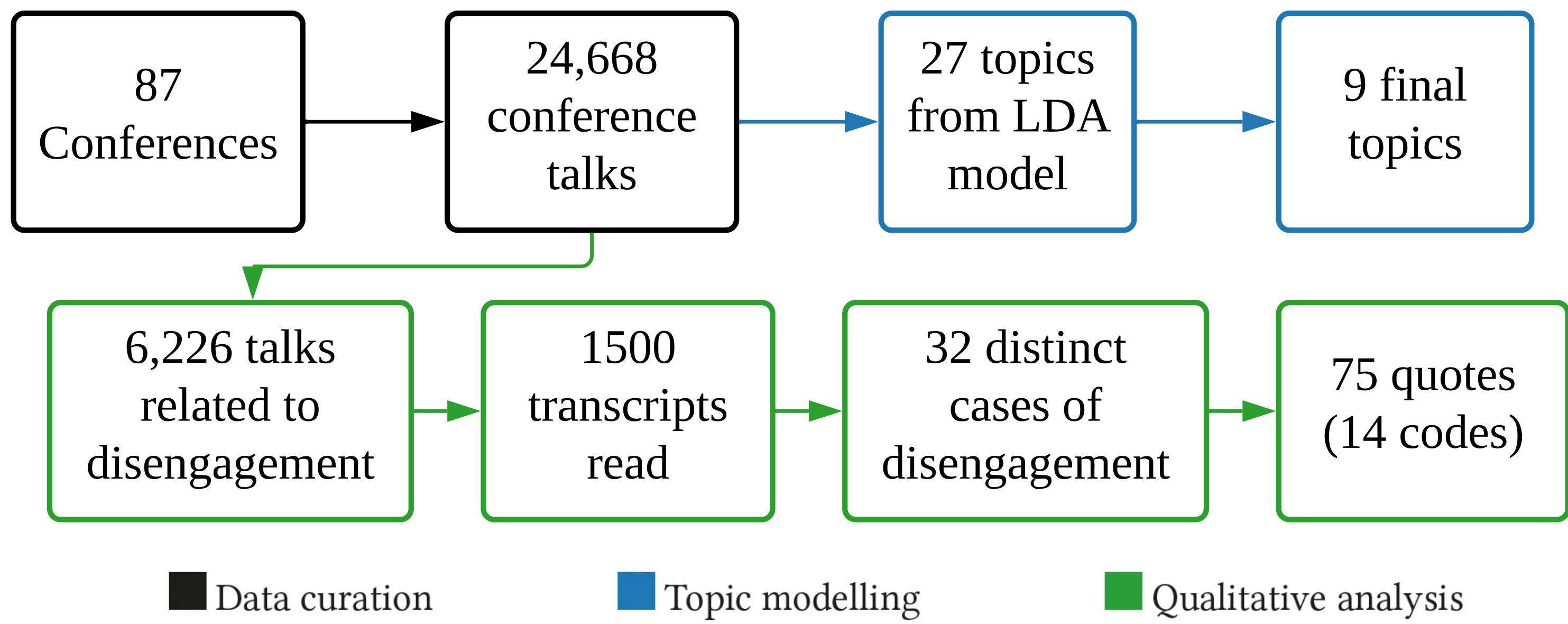
- Google search for top 30 links
  - “open source conference”; “open source conference call for proposals”

2. Collect metadata and **filtering** conferences:

- Two documented editions
- 50 attendees or > 10 speakers/talks

3. **Compile** data

- Generate separate text files for each video



### Applications

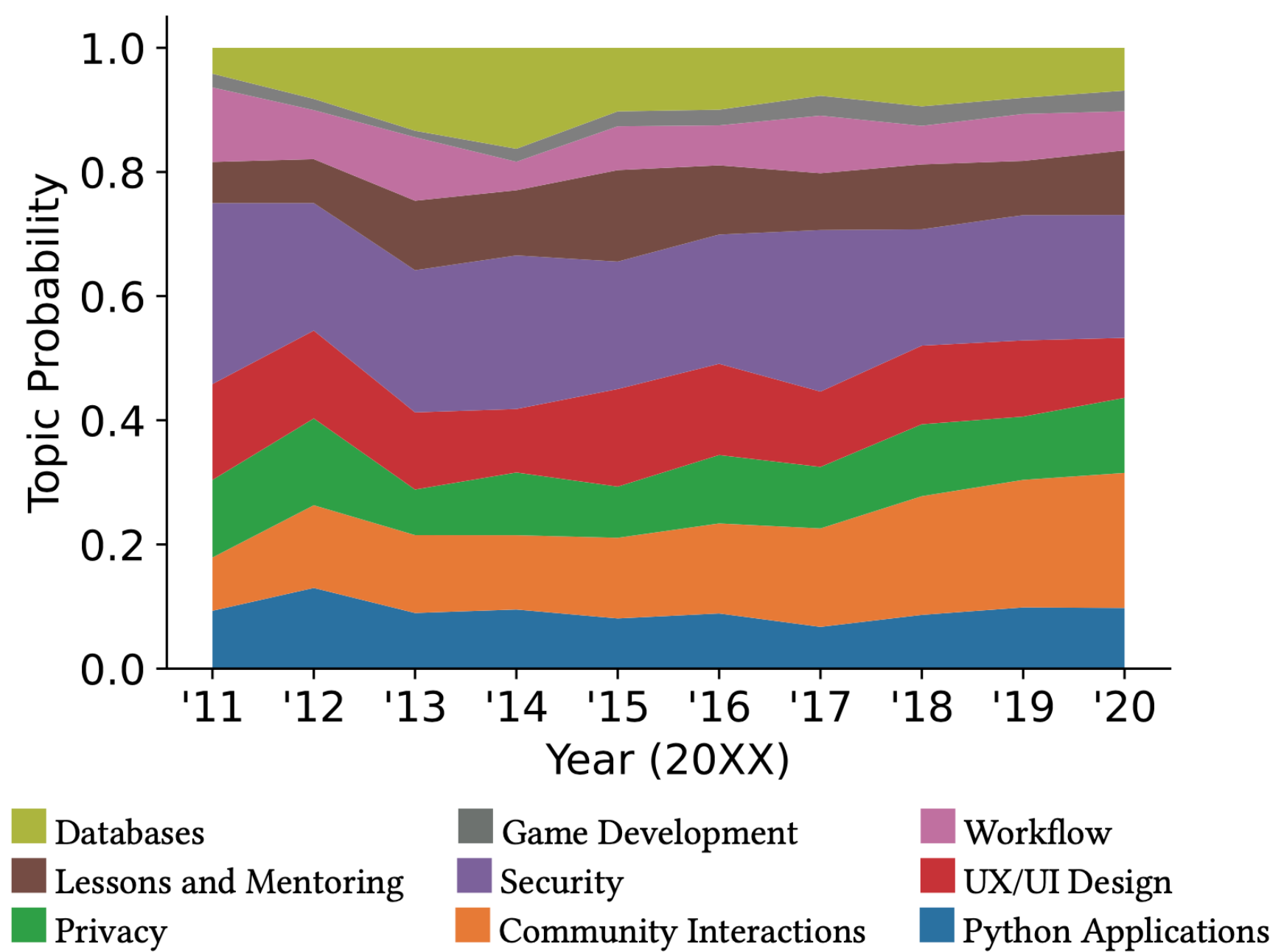
- Analyze the popularity and diffusion of tools and practices
- Understand common challenges discussed by practitioners
- Identify trends over time
- Identify major influences on certain communities
  - e.g., how the conference impacted the contributions for a certain project
- Filter to create samples by time, focus, size, etc.

### Topic Model

We see how the frequency of the most common topics in open-source talks have changed over 10 years.

#### Observations

- Community interactions is increasingly common.
- Talks about Databases have decreased after 2014.



### Disengagement Diaries

1. Generate a sample
  - Keywords: ‘leave’, ‘abandon’, ‘hostility’, etc.
2. Thematically code relevant transcripts.
3. Record reasons cited for disengagement, GitHub activity, and interventions

#### Cited Factors

- Cultural (32%)
  - Community leadership, policy disagreements, project direction, hostility
- External (18%)
  - Not enough time, self-doubt, health, left company, no longer useful
- Volunteering (50%)
  - Not enough time (internal), lack of support, no longer enjoyable, guilt, burnout

<https://disengagement-diaries.github.io>

Not Enough Time (Internal) No Longer Enjoyable

Who?

GitHub:

Project: Django

Why?

Anna cited three reasons for leaving their open source communities at DjangoCon US in 2018. These reasons arose because they had to change their priorities as new events unfolded in their life. The first issue was the lack of a work life balance as newer, more important responsibilities arose.

"People's living situations change or your priorities change and so people may say after a while I can't do it anymore for whatever reason." (15:44)

"Tech kind of takes over our whole life. I'm definitely guilty of that so I'm kind of like actively saying no." (30:00)

This led to a lack of time internally due to their shift in priorities. Although they still appreciated the Django community, Anna found contributing to open source to take away from more significant priorities, like a new job and changing living situations.

"I'm actually moving cities so I also kind of want to prioritize my personal life a little more I've served the community a lot and now it's kind of like time to look after myself a little more I'm out I also took on this

### For more information...

Read our paper at <https://doi.org/10.1145/3524842.3528488>  
e. truongkim@oregonstate.edu    /kimberly-le-truong-cs    /KimberlyTruong