CS 445/545 Final Report: An Analysis of Open Source Software Contributions

Alex Brelsford[1], Savannah Norem[2], Divyani Rao[3], and Ty Vaughan[4]

***Abstract*— In order to gain a better understanding of where open source software is coming from, we provide an analysis showing which locations are key sources for open source software, how these locations have contributed over time, and which countries are contributing the most to open source software development.**

I. INTRODUCTION AND MOTIVATION

Open source projects and communities have long existed. Over the past few years however, they have become an integral part of everyday life for those in the computer and data science communities. Because of this growth, we wanted to measure where these contributions were coming from, how active these places were in their contributions, and what companies were making them. Our motivation for this project was mainly how relevant this data is to our futures. As computer science students, knowledge of relevant open source projects and the ability to make valuable contributions to them will greatly benefit us in job searching.

II. DATA CONTEXT

One of the core aspects of our project is how well we can discover where to gather user and company information, how to gather the information, and how to represent the information with a format that can be easily interpreted. The following section discusses how user information was gathered and parsed for our project from our different sources: GitHub and StackOverflow/StackDump.

GitHub was our first choice for organization information gathering, but we quickly found that we weren't able to directly access the information we wanted. We discovered that when trying to access organizations linked to repositories, many repositories were marked as private, meaning that we were unable to access its information. Because of this, we had to change how we wanted to analyze organizational contributions. Our plan for analyzing which organizations contribute to open source became looking at the users related to repositories and obtaining their emails, from which we could pinpoint the specific organizations the users are affiliated with.

StackOverflow was another prime source of information, given that it is a publicly-available online forum where individuals can contribute to help others with software. However, it did pose a set of challenges for us to overcome in order to gather the information needed.

First, StackOverflow contains a rather diverse arrangement of question topics, some of which do not involve code. In order to justify using this information in our report, we need questions that directly involve code.

Second, the user information (email, location, and activity time stamp) needs to be accessed from all profiles queried. Regardless of which source we wanted to use to gather this information, the number of requests per 24-hour window was limited.

Third, the format of the location data was not consistent, and so we needed to parse out what we could. The methods for analyzing locations will be discussed in the following section.

To tackle the first challenge, we analyzed different questions from StackOverflow based on the coding language that is used within the question. Ten of the most popular coding languages were searched for: Python, C, C++, Java, JavaScript, R, Objective-C, PHP, SQL, and Ruby. If a question involved code from one of these languages, then we inferred that the users

[1]A. Brelsford is a CS student in the EECS College of Engineering at UTK. alexbrelsford at gmail.com

[2]S. Norem is a CS student in the EECS College of Engineering at UTK. snorem1 at vols.utk.edu

[3]D. Rao is a CS graduate student in the EECS College of Engineering at UTK. drao at vols.utk.edu

[4]T. Vaughan is a CS graduate student in the EECS College of Engineering at UTK. wvaugha2 at vols.utk.edu

involved with these questions were contributing to open source software because the code is now publicly available for anyone to use. Using this method of interpretation, we gain a representation of open-source contributions.

The second challenge posed a large bottleneck for our information gathering. To start, we had a few options for gathering user information:

- StackAPI provides an easy-to-use Python interface for gathering well-formatted information from StackOverflow. This issue with this resource is that for every 24-hour window, only 10 requests can be performed, getting a max of 100 results per request.
- Web scraping makes it harder to obtain the information we need, but allows about 600 user requests an hour if the requests are throttled properly to minimize blacklisting.
- StackDump offers a compact representation of user profiles that can be easily parsed. The downside is that it does not provide any contribution-activity information.

We started off by setting up a server job that ran a python script to perform requests every 24 hours using StackAPI. After a few days, we only had a couple hundred user profiles from questions. To make the most out of the available requests, we chose to use StackAPI to only query for question information. This allowed us to access about 10000 users, with the timestamp from when they were active, over the course of ten days.

The user profile links gathered in the process above were then used to scrape user locations off of profile webpages. Once we obtained the potential locations of all 10000 users gathered from StackAPI, we were then able to associate these locations with the time stamp information to visualize where contributions have come from over time.

Because so few users were gathered from this process, we also used the user profile dump provided by StackDump. This gave us access to over seven million user profiles. Over the course of a ten-hour window, we were able to successfully parse through almost two million of these users and output them into a neater format for location finding. Of these, a random sample of 25000 users was used to obtain non-time-based location data.

III. METHODS OF ANALYSIS

Given the data that we collected from webscraping StackOverflow, sampling from StackDump, and querying GitHub API, we needed a way to interpret the inconsistent location formats and a method for relating emails to organizations.

In order to parse out the locations from the inconsistent location formatting provided by different users, we performed the following steps to maximize our matches.

1) The location text was split using different characters: spaces, commas, and slashes. The end result was a list of potential words that pertained to a users country, city, or state if they live in the United States.
2) Different resources were gathered for location handling.
   - pycountry: this library contains all official country and nation-states along with their ISO Alpha-2 country codes, which are two-character mappings of each country.
   - MaxMind: this is a free database containing all cities in the world, the ISO Alpha-2 code for the country they are in, and the GPS coordinates of the city.
   - Manually created resources: we made our own listing of all states and their abbreviations, all countries and their geographical centers, all states and their geographical centers, and a mapping of all countries to the continent they are associated with.
3) The locations for each user were parsed out by searching the list of location information for names similar to all cities, states, and countries. Once all three location identifiers were searched for, the different resources were used to map countries to continents and locations to GPS coordinates. Coordinates were mapped in the following ways:
   - If a user only had a country identified, the GPS coordinates of the center were used to identify the location.
   - If a user only had a state identified, the GPS coordinates of the center of the state were used to identify location.
   - If a user only had a city identified, the most populous city with that name was used to identify the user, and that citys coordinates were used.

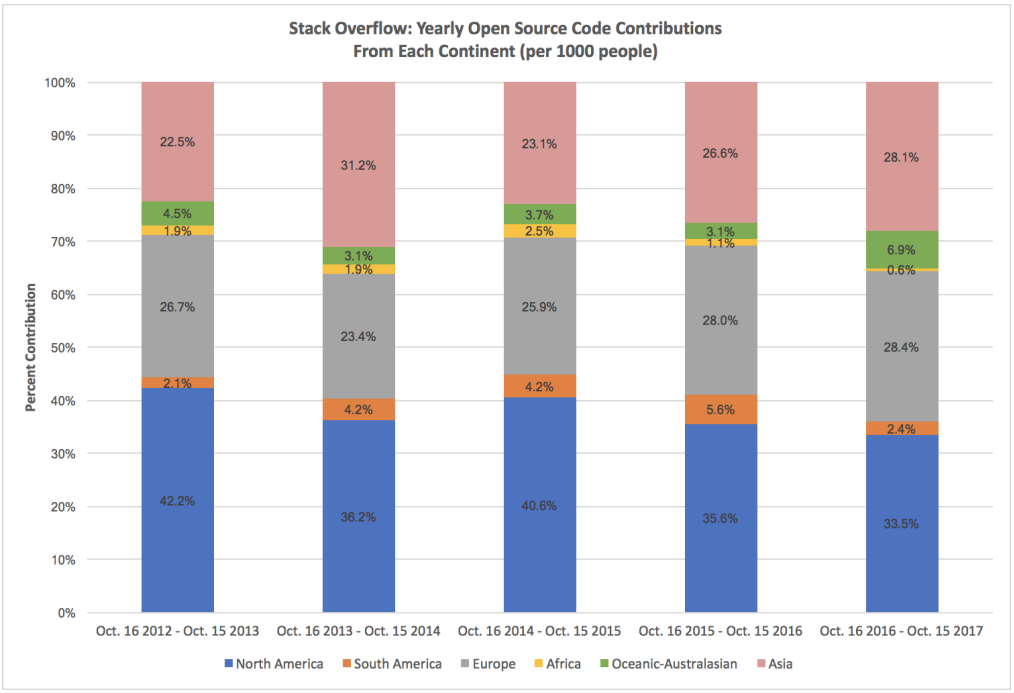

Fig. 1. Contributions from each continent per year.

The other form of analysis we needed to perform was linking organizations to emails. In order to match emails with organizations, we stripped all characters before and including the '@' symbol off of a sample of 10000 user emails that we had collected. This left us only with the email suffixes that represented either an organization or an email provider. We then filtered these to obtain all unique suffixes and how many of each was obtained. The goal of our analysis was to list all emails as belonging to an organization, possibly belonging to an organization, or not at all belonging to an organization.

To begin our analysis, we searched through all email extensions to search for '.edu', '.org', and '.net'. After manually checking a subset of these, we were able to confirm that the majority of emails with these suffixes pertained to an actual organization. Therefore, we used these extensions to identify organization emails.

A second search was done by performing text similarity analysis on email suffixes with corporation names obtained from lists such as the Fortune 500. If an email was more than 80 percent similar to a company name, then we marked it as belong to an organization.

Lastly, we also performed a manual analysis of emails, where each one of us on the team verified individual emails as belonging to an organization.

IV. RESULTS: WHICH LOCATIONS CONTRIBUTE TO OPEN SOURCE

Earlier in the report, it was mentioned that Stack Overflow data was separable into two groups: the group of users and locations that were gathered from question activity and thus are associated with a timestamp, and the group of users obtained from StackDump that are not associated with any activity. Thus, we can analyze the location of potential contributors based on activity and based on account creation. The results are listed in the section below.

A. *TIME-BASED CONTRIBUTION DATA*

First, the results of the locations based on user activity of 10,000 users is analyzed. The chart below sums up the activity of the last five years from each continent (excluding Antarctica from which no activity came).

As we can see in Figure 1 above, it appears that the majority of contributions are coming from North America for each of the last five years. Our sample also shows that other countries are contributing more over time. If we look at which countries from each continent are contributing the most over time, we get a consistent ranking of the following countries in order of most contributions to least (per continent). This is presented in Figure 2 below.

Most Significant Countries From Each Continent Per Yearly Contributions	
Continent	**Countries**
North America	
	1. United States
	2. Canada
South America	
	1. Brazil
Europe	
	1. Germany
	2. Great Britain
Africa	
	1. South Africa
	2. Egypt
Oceanic-Australasian	
	1. Australia
Asia	
	1. India
	2. China
	3. Russia

Fig. 2. Most active countries per continent.

B. NON-TIME-BASED CONTRIBUTION DATA

Out of the nearly two million accounts gathered from the StackDump StackOverflow user profile dump, about ten-hours were spent discovering locations for 25,000 randomly sampled users, and GPS coordinates were successfully identified for 23,909 of these users. These GPS coordinates are plotted below onto a world map. Because numerous users may map to a single GPS coordinate, some noise was added to the coordinates that had many users so that they stand out. The results are in Figure 3.

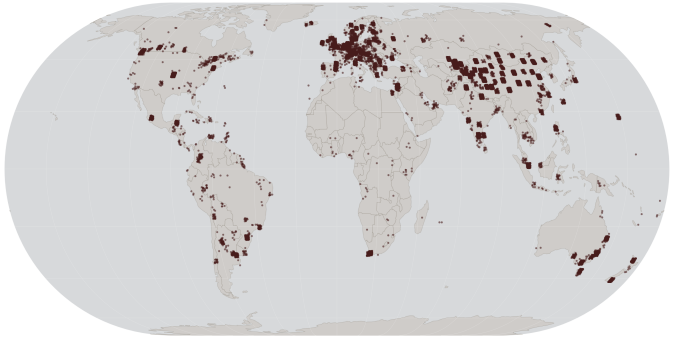

Fig. 3. Identified contributing locations from around the globe [3].

Looking at Figure 3, the results appear much more widespread in the Eastern Hemisphere than they do in the Western Hemisphere. A big part of this is due to how if a city or state isnt found for a user, that users GPS coordinates are marked as the center of their affiliated country. Because there are fewer countries in the Eastern Hemisphere, it appears much less widespread. The more numerous countries in Europe and Asia cause there to be more widespread identified locations.

In order to get a better idea of where open source contributions are coming from, the continents are listed in Figure 4 from having the most to least contributing users, along with the three countries from each continent that have the most contributors.

Top Contributors Based on Profile Location			
Continent	**Profiles per Continent**	**Countries**	**% Continental Contribution**
1. North America	14009		
		1. United States	93.2%
		2. Canada	6.0%
		3. Mexico	0.5%
2. Europe	6170		
		1. Great Britain	23.4%
		2. Germany	10.9%
		3. Sweden	9.4%
3. Asia	1575		
		1. India	36.8%
		2. Israel	17.0%
		3. China	7.0%
4. Oceanic-Australasian	1491		
		1. Australia	84.8%
		2. New Zealand	15.0%
		3. Fiji	0.1%
5. South America	428		
		1. Brazil	55.6%
		2. Argentina	28.7%
		3. Columbia	6.1%
6. Africa	236		
		1. South Africa	76.7%
		2. Egypt	10.2%
		3. Kenya	3.4%
Total: 23909			

Fig. 4. Countries per continent with the most contributors.

The data in Figure 4 shows that the above claim about user location density in Figure 3 is valid: most of the users are identified only by countries, and thus it appears that fewer users in North America contributed compared to Europe or Asia even though North America contains the most contributors. If we want more specific information, it is helpful to see which identified cities are contributing the most to open source. The significant cities per continent are listed in Figure 5.

While observing the data in Figure 5, it is important to keep in mind that not all users specified their city, meaning both that not all cities that should be represented are and that cities may not be represented

Top Contributing Cities per Continent			
Continent	Country	City	Number of Contributors
North America			
	1. Canada	Vancouver	174
	2. Canada	Calgary	83
	3. Canada	Ottowa	95
Europe			
	1. Sweden	Stockholm	163
	2. Norway	Oslow	111
	3. Germany	Munich	76
	4. Denmark	Copenhagen	68
	5. Great Britain	Edinburgh	62
Asia			
	1. India	Bangalore	161
	2. Turkey	Istanbul	62
	3. India	Chennai	55
	4. Japan	Tokyo	51
	5. India	Mumbai	45
Oceanic-Australasian			
	1. Australia	Sydney	305
	2. Australia	Melbourne	220
	3. Australia	Brisbane	127
	4. New Zealand	Auckland	84
	5. Australia	Perth	81
South America			
	1. Argentina	Buenos Aires	60
	2. Brazil	Rio De Janeiro	32
Africa			
	1. South Africa	Cape Town	65

Fig. 5. Cities per continent with the most contributors.

evenly. However, Figure 5 does offer a glimpse of which cities may be the top contributors.

One last specifier for location that can be observed is the state that a user within the United States lives in. The top contributing states identified within our sample of user locations are listed in Figure 6.

Top Ten Contributing States	
State	Number of Contributors
1. California	2065
2. Washington	880
3. New York	801
4. Texas	746
5. Massachussetts	468
6. Illinois	387
7. Pennsylvania	378
8. Florida	347
9. Ohio	317
10. Colorado	312

Fig. 6. States with the most contributors.

Similar to how city locations were not specified by all users, the same applies to states. Not all individuals within the United States listed a state of residence, and thus the states may not be represented with perfect accuracy, though the results do still provide information as to what can be expected.

V. RESULTS: WHICH ORGANIZATIONS CONTRIBUTE TO OPEN SOURCE

The results were collected from the email addresses that we received from the GitHub API. Then, we made a list of all the email addresses with the same domain names. Since there was no other way of separating the email addresses as companies and organizations versus individuals, we split the data into four sets and then manually categorized them as the following:

- Yes: This email address represents a company or an organization.
- May be: This email address might be a company or an organization.
- No: This email address does not represent a company or an organization.

A domain name with more than one instance of email addresses is considered an important email in our results. The pie chart in Figure 7 below shows the percentages of Yes, May be and No for all of the emails that we collected, and the pie chart in Figure 8 below shows the percentages for just the important emails.

There were multiple instances of important emails in our data. When we observe the two charts below, we can see that results from important emails have a higher tendency to be an organization. A significantly higher percentage of important emails were categorized as a company in comparison with the results from all emails. To to be precise, the difference between the percentages for ”Yes” in important emails verses ”Yes” over all emails was 44 percent. Another observation from this chart is that most of the emails that we obtained from GitHub did not correspond to a company or an organization.

VI. RELATED WORK, LIMITATIONS, AND ISSUES

The section below discusses prior work that is related to our project, followed by a discussion of some of the limitations and challenges encountered during our project.

A. RELATED WORK

A measurement of the languages being used on StackOverflow showed first that the majority (63.7 percent) of their traffic came from ”high-income” countries. It went on to show that Python and R are visited

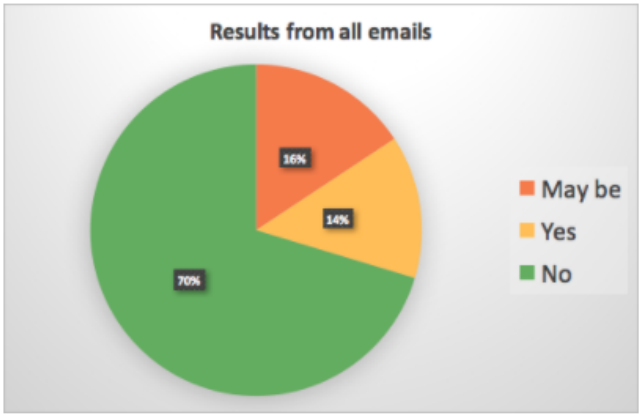

Fig. 7. Results from all emails.

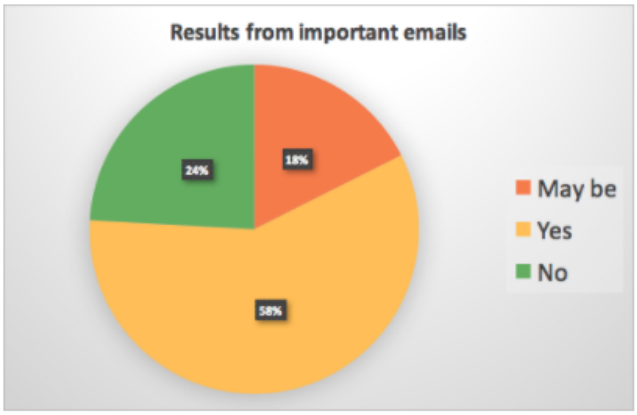

Fig. 8. Results from important emails.

twice and thrice as frequently in high-income countries as they are in the rest of the world, respectively. This discrepency in popularity only grew when looking at scientific packages for these languages, such as numpy or ggplot2 [1].

A 2017 open source job report found that 89 percent of hiring managers found it difficult to find applicants with talent at developing open source projects. The report also found that 47 percent of employers were willing to pay for professional certifications, and were looking for applicants with expertise in Cloud, Web Technologies and Linux. The hiring managers were looking for people to fill positions titled ”Developers”, ”DevOps”, and ”SysAdmins”. The report found that, as a whole, employers are more willing to pay for traing for employees in open source technologies [2].

B. LIMITATIONS AND ISSUES

One of the major limitations of this project was the emails. We initially thought that classifying whether a person contributing to a project was with a company or not could be relatively easily determined by their email address. Once we pulled the email addresses however, we realized that this was close to impossible to do accurately. As a group, we decided that web addresses that seemed to be people’s names were probably not organizations. We also decided that ’.net’ and ’.org’ web addresses probably were organizations. Taking this into account, there is still the possibility that people who work for major companies were given a no because they used their personal email address. There was still a lot of gray area however, especially web addresses that did not seem to be English.

Another limitation is the difficulty we had defining what open source means. While there are some open source projects that are large and well known, there is a large culture of people using open source communities to get help on personal projects.

We also had an issue with the way we classified

location data. For instance if the way we designed the location search favored the United States since we assumed that most of the users would be from here. For instance if a user only listed that they were in Paris, without any country attached to it, our search would place them in the first alphabetical state with a city named Paris. If there was a city with the same name in the state of Georgia and the country Georgia, our search would have placed them in the United States as well. We think that these misclassifications happened statistically rarely since the vast majority of users had a country listed, and we dont believe that city overlap between the two Georgias is significant either.

VII. FUTURE WORK AND CONCLUSIONS

Going through emails we noticed that it was difficult to determine if an email was related to an organization or not. To better classify whether not an email is related to an organization or not we would need to do more in depth searching. To tie in the other part of our project we wanted to get location data on the organizations so we could see where the tech hotspots are for open source software contributors. On top of that we wanted to do a broader search over multiple open source sites rather than just GitHub. This would enable us to see if organizations are contributing to other open source sites more often than GitHub.

As for location data, we wanted to create a better classification system so we could avoid misplacing a users location. Since the sites we analyzed are predominantly in English, there is a good chance this influenced the fact that most users we found were in America and Western Europe. Our user location data came from a single source which has a large user base and can give a good view of where open source is coming from, but we want to see if our data is consistent when coming from other sites, especially ones that are not in English.

The biggest problem we faced with this project was getting to and gathering enough data. We had issues with people not using valid locations and would leave out the country in some cases. We ran into issues when we noticed GitHub organizations were inaccurate and the emails werent always accurate. Once we figured out ways around these issues we were able to start fully analyzing the data. Most of the data represented what we expected. Most open source is coming from developed countries, and the majority of it is coming from North America. What we didnt expect was how the emails turned out. Most emails belonged to a single user and would only show up once and the ones that showed up multiple times were mostly organizations. Emails that showed up more than once were very obvious organizations since they belonged to large companies like Google, Redhat, and Microsoft.

REFERENCES

[1] M. Byrne. Coders in wealthy and developing countries lean on different programming languages. Online, 2017.

[2] Steven Vaughan-Nichols. Open source professionals are more in demand than ever: Dice and the linux foundation's 2017 open source jobs report reveals linux and open-source jobs are hotter than ever. Online, 2017.

[3] Derek Watkins. Geo point plotter. Online, 2017.