Open Source Contributions

CS 545

Members: Savannah Norem, Divyani Rao, Alex Brelsford, Ty Vaughan

Date: Nov. 29, 2017

Introduction and Motivation

- Open source software is free, easily available, and may be redistributed and modified
- What is the relationship between open source and where it comes from?
- What countries are contributing to open source software?
- What organizations are contributing to open source software?

Location Data - Context

- The data for answering where open source has come from over time and the locations of who contributes are gathered separately.
- Data for Contributor Locations Over Time
  - This data is collected using StackAPI and StackOverflow. StackOverflow questions were searched for code in Python, R, SQL, C, C++, Java, JavaScript, Ruby, HTML, and CSS. The users involved with these questions were flagged and their locations were grabbed.
  - A total of 10,000 users were obtained.

Location Data - Context (Cont.)

- Data for Contributor Locations
  a. This data was collected from StackDump's user dump for StackOverflow.
  b. A random sample of 25000 users was obtained.

- The format of user profile locations was not consistent between profiles. Different parameter separators are used (commas, slashes, semicolons).
- The location data contains misspellings and fake locations.
- Locations have any combination of a country, city, and/or state. For this project, we wanted to parse out as best as we could a specific location pertaining to a contributor.

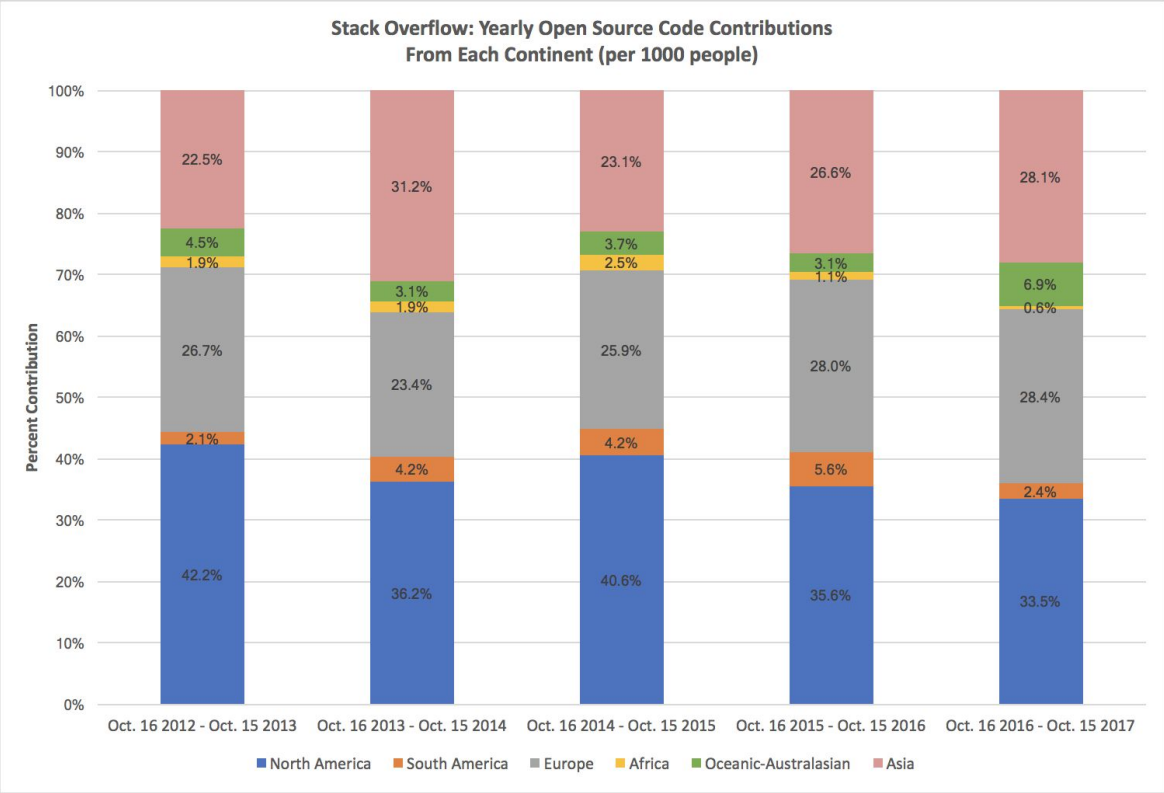
Location Data - Analysis

- All locations were separated by commas, slashes, or other symbols that appeared as designators for splitting up location fields.
- Multiple resources were used to gather location information to compare user profile locations to:
  - Pycountry: allowed us to map countries to their two-character ISO tag
  - MaxMind world city database: allowed us to map cities to GPS coordinates
  - Self-generated lists connected states to GPS coordinates and countries to both continents and GPS coordinates.

Location Data - Analysis

- Each element of a user's location was searched for in our list of countries, cities and states.
- As matches were obtained, specific GPS coordinates were identified based on the center of either the city, state, or country identified (in that order)

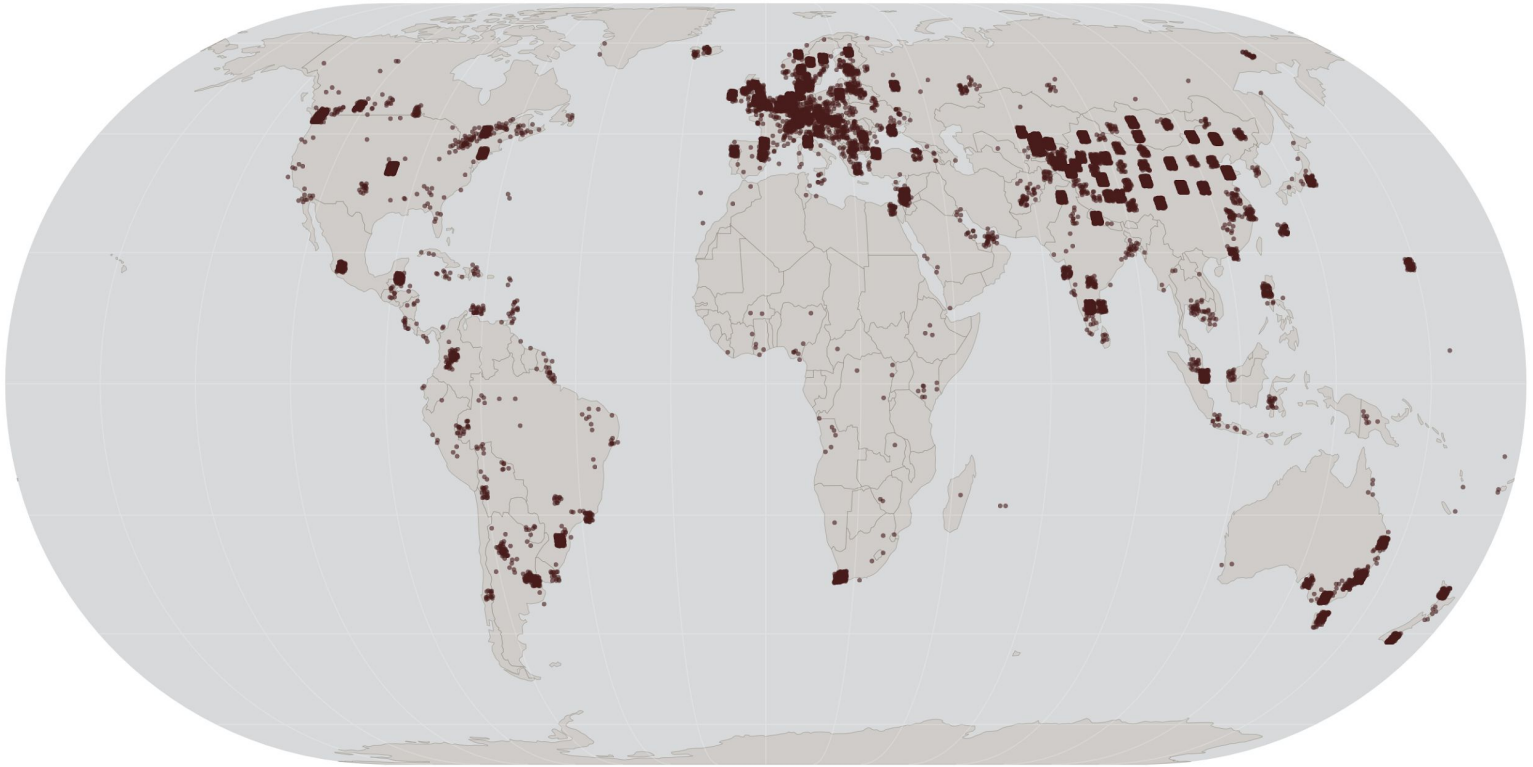
Location Data - Time-based Results



Location Data - Time-based Results (Cont.)

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

Location Data - Non-Time Based Results



Location Data - Non-Time Based Results (Cont.)

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			

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

Location Data - Non-Time Based Results (Cont.)

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

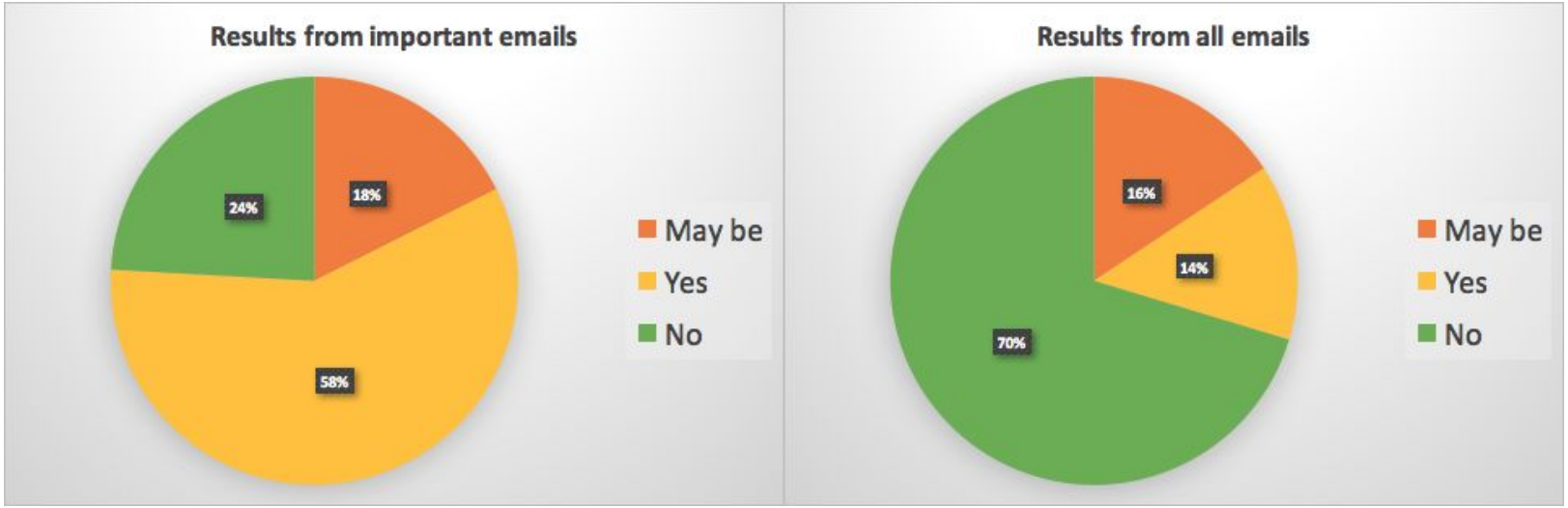
Company Data - Context

- The data for answering who contributes to open source comes was captured over multiple days
- Emails were obtained through the GitHub API
    - We acquired 10,000 emails and then grouped them by suffix
- Grouped emails were then labeled as an organization, not an organization, and maybe an organization

Company Data - Context (Cont.)

- Not all the emails were accurate
  - Not actual email addresses
  - Mistyped email addresses

Company Data - Results



Company Data - Analysis

- Important emails showed up multiple times in our list
- Results from important emails had a higher tendency to be an organization
- Results from all emails show that a majority of public emails were not related to an organization

Related Work

- *2017 Open Source Jobs Survey and Report*:
  - Showed that further developed and wealthier countries had more contributors to open-source software
  - Also showed that larger companies, such as Microsoft, AT&T, IBM, and RedHat, contribute much more to open-source than smaller, less-wealthy companies
- *Stack Overflow: David Robinson's Report*:
  - Showed that different languages are preferred in different locations around the world

Limitations and Issues

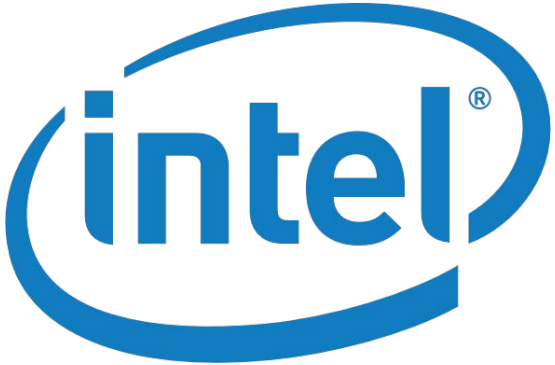
- Location classification algorithm has the potential to misclassify when city names are replicated and a state / country name is not listed
- The locations were first placed in the United States since we assumed that most contributions are from the United States
- Categorizing companies and organization vs. individuals solely by email addresses
- Determining what counts as an open source project

Future Work

- Find out locations of companies
- Gather more data from other sources
- Better classification of location data
- Better classification of company data

Conclusions

- Most of the location data shows open source is coming from developed nations such as United States, Canada, Great Britain, Germany and India.
- Most of the open source organizations were from Yahoo, Red hat and Intel.



Thank You!