Does the OpenStreetMap Forum reflect a crowd or community model of participation?

Jamie Fawcett

February 2019

Word count: 2491

The following paper is a full public reproduction of an assignment produced by the author for the 'Accessing Research Data from the Social Web' course, in partial fulfilment of the requirements for the degree of MSc in Social Science of the Internet at the Oxford Internet Institute at the University of Oxford.

All code used to produce the analysis and figures in this paper is available on GitHub under an open licence.

Does the OpenStreetMap Forum reflect a crowd or community model of participation?

Introduction

OpenStreetMap (OSM) is a global project which aims to create a free, openly licensed database of geographic information through mass collaboration. It is an example of 'social production of knowledge', similar to Wikipedia or Open Source Software (OSS) (Benkler, 2006). Haythornthwaite (2009) argues that all social production projects exist on a spectrum between crowd and community models. Crowd-based models rely primarily on large numbers of lightweight or casual contributors, whereas community models depend on smaller numbers of heavyweight or serious contributors. Understanding where a particular project sits on this spectrum can have significant impact on the way they design systems and allocate limited resources (Kittur, Chi, Pendleton, Suh, & Mytkowicz, 2007).

Previous computational research has highlighted the reliance of a wide range of successful projects on heavyweight community contributors, including in OSS development (Chełkowski, Gloor, & Jemielniak, 2016), contributions to Wikipedia (Ortega, Gonzalez-Barahona, & Robles, 2008) and contributions to the OSM database (Ma, Sandberg, & Jiang, 2015). However, one aspect which has typically been ignored is the multifaceted nature of contribution to these types of projects. Beyond contributing data, OSM contributors also develop the open source codebase and tools, maintain its wiki and communicate with others through its plethora of different communication channels.[1]

In this paper, I examine one of OSM's communication channels, the OSM forum,[2] to understand whether it reflects a crowd or community model of participation. To do this I identify the distribution of contributors by level of contribution to determine whether lightweight or heavyweight contributors are responsible for producing the majority of contributions. To begin, I layout the methodological approach with a particular focus on the methods used to collect the relevant data. Next I present the results of my analysis in characterising the distribution of contributions and finally I discuss the implications of these results to understanding the participation model of the OSM forum.

[1] https://wiki.openstreetmap.org/wiki/Contact_channels
[2] https://forum.openstreetmap.org/

Methodology

Determining whether participation in the forum is indicative of a crowd or community model requires data about the number of contributions by each contributor. As the OSM Forum does not have public user profile pages which list individual number of contributions, I had to collect data on individual contributions (forum posts) and then aggregate this data by contributor.

With this in mind, I chose to collect data from the entire forum, rather than employing a random or stratified sampling approach at the sub-forum level. While these sampling approaches are feasible, collecting data at the post level creates potential for mischaracterization of contributors. Namely, if individual contributors post primarily to single sub-forums but also occasionally outside, then any sampling approach might misidentify them as heavyweight or lightweight depending on which sub-forums are sampled. The likelihood of this is potentially increased by the fact the sub-forums are organised both geographically (individual countries) and thematically (wiki maintenance), and the amount of content in each set varies dramatically.

I also highlight the ethical considerations of collecting and analysing the data. Although participants post on the forum with the expectation that it will be available on the public web, they do so primarily in order to communicate with others OSM users and may not have considered the potential for analysis, leading to issues of consent. In addition, to understand contributions by contributor it is necessary to capture participant pseudonyms. These directly identify participants and may be used in conjunction with other sources, for example the OSM database itself, to learn things about the individual. With these considerations in mind, I treat the data collected as personal data keeping only a local copy on a password protected laptop with encrypted hard drive, not sharing the data and ensuring it is deleted at the end of the course.

Data collection

To access and process data from the forum, I used an Anaconda distribution of Python 3 within a Jupyter notebook environment.[3] The data collection approach I used was determined by the structure and format of the forum itself, an overview of which is presented in Figure 1. The available formats lead me to use a combination of HTML

[3] Version: Anaconda: 4.6.1, Python: 3.6.5, Jupyter: 4.4.0. See GitHub for code.

scraping and RSS feeds to extract data from the forum. As there was no index of individual posts, I chose to iteratively navigate through the hierarchical forum structure from the index page.

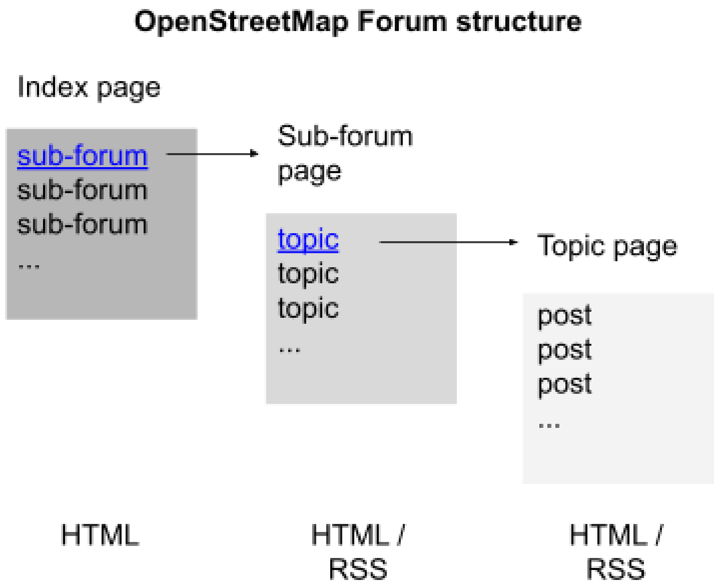

Figure 1. Structure of the OSM Forum.

To do this, I defined a set of nested functions which collected iterated through all items at each of the three levels and stored the appropriate data. Within these functions, I used the Python library requests[4] to access webpages, BeautifulSoup4[5] to parse HTML, feedparser[6] to parse RSS feeds and Pandas[7] to store data. Table 1 outlines the broad functionality of each of the nested functions used.

Function	**Process outline**	**Input**	**Output**
getForums	Iterate through sub-forums listed on index page and call *getAllPages*.	Index url, filename	Write entire sub-forum DataFrame to csv
getAllPages	Iterate through all sub-forum pages and call *getTopics* on each page.	Sub-forum id	Return DataFrame of all topics in sub-forum.
getTopics	Read sub-forum page Iterate through all topics and call *getPosts* on each topic.	Topic page	Return DataFrame of all topics on page.
getPosts	Read topic RSS feed Iterate through all posts and store data.	Topic id	Return DataFrame of all posts.

[4] https://github.com/kennethreitz/requests
[5] https://pypi.org/project/beautifulsoup4/
[6] https://github.com/kurtmckee/feedparser
[7] https://pandas.pydata.org/index.html

Table 1. Structure and functionality of data collection code.

Instead of running through the process in depth, I highlight some of the key design decisions and limitations in this approach.

1) Iterative approach to data collection and storage

I decided to store the data collected in Pandas DataFrame objects in order to limit the amount of post-collection processing required for analysis. These objects were created iteratively at each stage by first iteratively creating lists of smaller DataFrames and then concatenating them after completion of all iterations. For topic and sub-forum provenance data (e.g. topic url), I assigned this data to all relevant records after iterations were complete. Given the potential amount of data in sub-forums and the complexity of collection, I iteratively append data from each sub-forum to a csv file rather than after the entire process to run.

2) Creating unique id and limiting unnecessary low-level data collection

To ensure that records are not merged when the DataFrames are concatenated at higher levels, I assigned unique ids based on topic_id and count value of the enumerate function. As we only require author_id to generate the analysis of number of contributions, only this and a timestamp are collected for each post. Even though OSM does not allow display name changes (author_id), I initially considered using an available author_uri however exploratory analysis revealed no cases of where the author_uri differed by author_id but several cases where author_uri was unavailable.

3) Approaching multiple sub-forum pages

As sub-forums can have multiple pages, I defined a function (getAllPages) which used the HTML text elements on the first page to identify the maximum number of pages in each sub-forum as this element is always visible. I then used the range function to iterate through each page from 1 to maximum by manipulating the URL string. I deemed it necessary to take this approach over the use of a generator as all page numbers above the maximum page number when passed to URL return the first page of the sub-forum.

4) Error logging

Throughout the process of data collection there were a number of different errors, so I created a separate error log file to record errors and passed over them. Examples of the errors produced included where sub-forums or topics had no accessible contents, in some cases due to an access error. Although this allowed the script to collect all data, initial data analysis revealed uncaught errors or incorrect error handling lead to gaps in data.

5) RSS feed limitations

Because RSS feeds only return the most recent 15 entries for each topic, after initial data collection I had to pass a list of all topics with more than 15 entries (as recorded in reply count) through another set of two nested functions. In these functions, I adopted an HTML scraping process to collect data on posts, iterating through multiple pages using the same approach as getAllPages. This process produced an additional DataFrame which was saved to CSV.

Data preparation and analysis methodology

Once all the data had been collected I prepared the data for analysis. Because of the append method used to store sub-forums the column titles of all but the first DataFrame were stored as individual rows, so I removed these. The data from the original RSS-based process and that obtained through HTML scraping were concatenated into a single DataFrame and checked for duplicates. In order to analyse contributors by contributions, I created a new dataframe using the value counts function to get a Series object with each contributor and their number of contributions.

To answer the research question of whether heavyweight or lightweight users provide greater contributions to the OSM forum, I begin by using simple descriptive statistics and a histogram plot (using the Seaborn library) to analyse the distribution of contributors by contribution. To understand whether this distribution may have power law properties, I plot contributors and contributions on a log-log scale. In order to characterise the scaling of this distribution, I follow the approach of Ma et al (2015) in using a head-tail breaks. This involves recursively analysing the head of the contributor distribution (where number of contributions are greater than mean number of contributions) while the percentage of contributors in the head is under 40%. The number of times the data can be split plus 1 gives the ht-index which explains “how many times the scaling pattern of far more small things than large ones recurs in the data”. Finally, I calculate the percentage of contributions accounted for by the top X%.

Limitations and challenges

There are three key limitations and challenges in the approach taken. Firstly, for the first six months the forum required users to create new accounts or use guest accounts rather than use their main OSM account.[8] While users can no longer post with guest accounts, they are still able to establish separate identities. Both these issues may affect the interpretability of

[8] https://wiki.openstreetmap.org/wiki/Forum

the data if users use more than one account, for example giving the impression of more lightweight contributors than there are. Secondly, a small number of topic details were not stored in the output data arising from an unknown error, while this does not affect overall analysis it may result in the HTML scraping of posts not accounting for when these topics were over 15 posts long. As such, in future studies it might be preferable to integrate the HTML post scraping directly into the initial data collection approach. Finally, there may be a number of limitations in the error logging approach and in the efficiency of the collection process which would need to be addressed in future studies.

Results

Using the data collection method laid out above, I captured 710,445 unique records of forum posts containing author name, timestamp as well as provenance details about the topic and sub-forum. Table 2 provides a comparison between the statistics available on the forum index page and the data collected. As the table shows, only a relatively small percentage of posts (0.3%) and topics (2.92%) were not collected. The number of missing topics may be an overestimate as some topic details were not stored in the data outputted and have not been addressed as they are not required for the purpose of analysis.

	Forum stats	**Data captured**	**Missed**
Total sub-Forums	88	88	0% (0)
Total topics	42,699	41,452	2.92% (1247)
Total posts	712,862	710,445	0.33% (2417)

Table 2. Comparison between OSM Forum statistics and data collected.

Analysis

As displayed in figure 2, the distribution of contributors by number of contributions is very heavily skewed. In order to visualise the distribution meaningfully I used a logarithmic scale for the y-axis (number of contributors). As shown in the figure, the vast majority of contributors make very few contributions. This is further evidenced in the descriptive statistics in table 3 where the mean number of contributions is 39.3 but the median is 3.

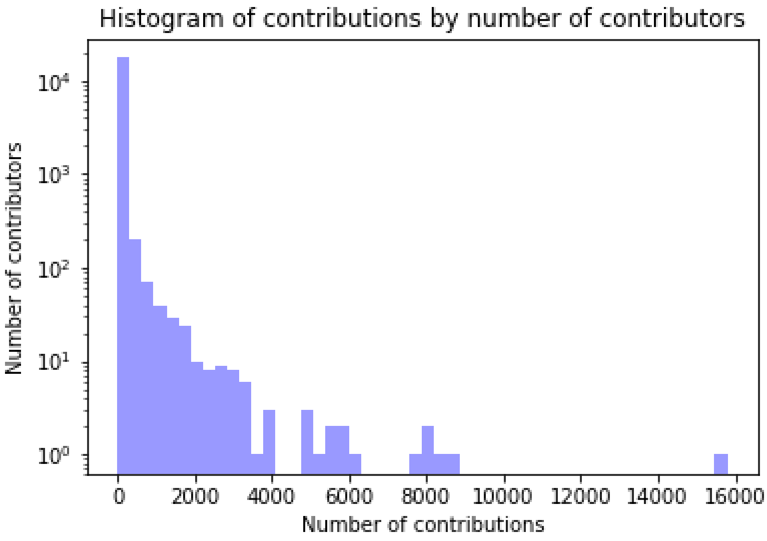

Figure 2. Histogram of contributions to forum by number of contributors (on log scale)

	count	mean	std	min	25%	50%	75%	max
author	18090	39.27	284.13	1	1	3	7	15795

Table 3. Descriptive statistics of contributor's contributions

Given this distribution, I present a log-log plot of contributors by contribution in figure 3. The plot suggests an approximately linear relationship with a significant head which is an indicative characteristic of a power-law distribution.

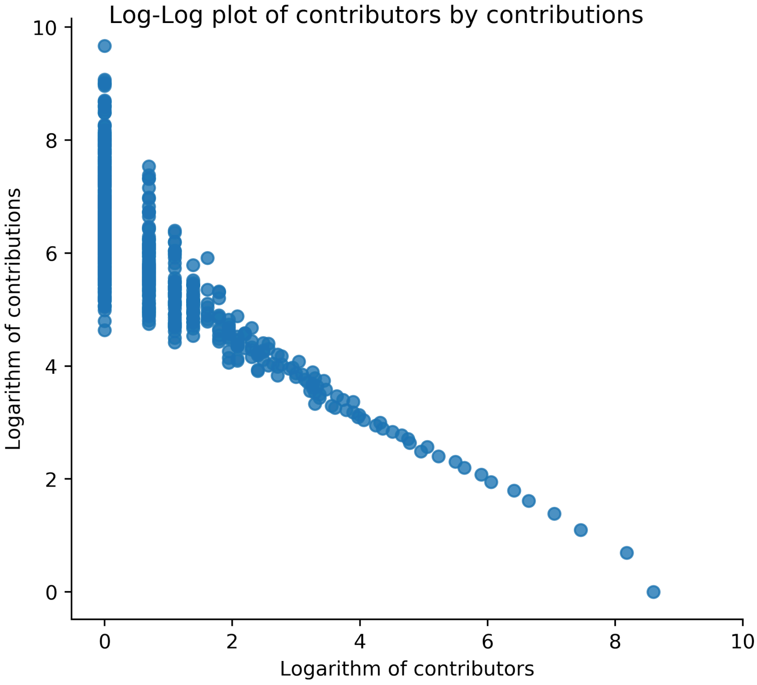

Figure 3. Plot of log contributors against log contributions

To characterise the scaling properties of this distribution, I carry out head-tail breaks analysis which gives an ht-index of 6. From Table 4, we can observe that over 90% of users contribute less than the mean number of contributions. From Table 5, we can observe that the top 10% of contributors provide just under 90% of total contributions while the top 1% contribute over 50%.

Break	Total contributors	Mean contributions	Contributors in head		Contributors in tail	
			Count	Percentage	Count	Percentage
1	18090	39.27	1651	9.17	16439	90.87
2	1651	383.20	360	21.80	1291	78.20
3	360	1323.93	110	30.56	250	69.44
4	110	2805.29	35	31.82	75	68.18
5	35	4861	13	37.14	22	62.86

Table 4. Results of head-tail break analysis

	Percentage of all contributions
Top 10%	89.90
Top 5%	82.45
Top 1%	53.39

Table 5. Percentage of all contributions carried out by top X% of contributors.

Discussion

The distribution of contributors, descriptive statistics and the log-log plot all appear to mirror findings presented in previous research around the distribution of direct contributions to the OSM database. With a scaling factor (ht-index) of 6 it is similar to the distribution found by Ma et al (2015) for direct contributions to the OSM database (ht index = 7). With just over 90% of contributors providing less than the mean number (39) of contributions the forum appears to be strongly centred on a community model of participation. Providing further evidence of this is the finding that the top 10% of users account for almost 90% of total contributions.

Although the findings of this initial research favour a community-based model of participation, further research should be conducted to examine the true nature of this participation. For example, by using the data collected to construct a co-contribution network of contributors by topic. Examining the topography of this network would likely provide a more detailed analysis of the structure of crowd vs community dynamics. Additional analysis of the data by type of contribution (original post vs reply) could also provide insight into the nature of lightweight vs heavyweight contributions. Finally, conducting similar analyses of the other communication channels could provide greater insight into the collaboration dynamics of the OSM project as a whole.

Bibliography

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom Contract : Freedom in the Commons*. New Haven, US: Yale University Press.

Chełkowski, T., Gloor, P., & Jemielniak, D. (2016). Inequalities in Open Source Software Development: Analysis of Contributor's Commits in Apache Software Foundation Projects. *PLOS ONE*, *11*(4), e0152976. https://doi.org/10.1371/journal.pone.0152976

Haythornthwaite, C. (2009). Crowds and Communities: Light and Heavyweight Models of Peer Production. In *2009 42nd Hawaii International Conference on System Sciences* (pp. 1–10). Waikoloa, Hawaii, USA: IEEE. https://doi.org/10.1109/HICSS.2009.137

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. *World Wide Web*, *1*(2), 19–29.

Ma, D., Sandberg, M., & Jiang, B. (2015). Characterizing the Heterogeneity of the OpenStreetMap Data and Community. *ISPRS International Journal of Geo-Information*, *4*(2), 535–550. https://doi.org/10.3390/ijgi4020535

Ortega, F., Gonzalez-Barahona, J. M., & Robles, G. (2008). On the Inequality of Contributions to Wikipedia. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)* (pp. 304–304). Waikoloa, HI, USA: IEEE. https://doi.org/10.1109/HICSS.2008.333