CS 2704
Term Project Proposal

Matthew Kenneth Peterson
(3719754)

November 14, 2023

1 Background

Of all aspects of the culture surrounding computer science, and software development in particular, the concept of open-source software and its development is perhaps one of the crowning achievements of the community at large. It is a remarkable tool no matter its shape or size, with applications in education, encouraging greater collaboration and more, all with an emphasis of accessibility. An open-source program is, by definition, open to *anyone* to modify and use, so long as its license permits it. In turn, this means that the majority of open-source software is free for anyone to use, regardless of whether or not they intend on editing the source code.

It seems as though many developers who dedicate their time and effort to open-source endeavours do it not only for possible personal gain, but for the love of their community. Participation in open-source software projects for most programmers means doing the work on their own time, away from a professional environment. Further, because a majority of open-source projects are made available for free, this means that most who contribute to these projects tend to see little if any monetary gain from their endeavours aside from acts of charity (i.e. donations).

In this way, it is sufficient to say that open-source contribution is much akin to volunteer work. Hard work with nothing expected in return, giving your time to some sort of cause. But that is simply by definition, and I am curious to see if it holds up in real world practice. Therefore, my goal for this project is to see if there is a possible correlation between areas of the world where it is more common for individuals to participate in bettering their communities, and areas of the world where it is more common for software programmers to contribute towards open-source endeavours. Currently, my hypothesis is that there will be a strong correlation between the two.

2 The Datasets

Unfortunately, this is not a heavily surveyed topic. Metadata for these sorts of studies is difficult to gather, especially when necessary APIs are not made available by the major open-source software (OSS) platforms. However, what data there is available on the topic may reveal some interesting results.

For this exploratory project, I will be using the data collected by a group of researchers out of Vienna from 2021, in which they plotted the density of GitHub users who made over 100 contributions to open-source software, by country and certain sub-regions. [1] The reason they had done this was to attempt at establishing a positive correlation between a country's gross national income (GNI) and its open-source contribution count. It is remarkable work, as they have gone so far as to account for discrepancies in the number of accounts against the number of actual individual persons who have used the service. They have done this by taking the GitHub archived history of all public commits, and queried its saved email and personal information against the GitHub account lookup API. Considering the tools at their disposal, this may be the best snapshot we have into the location data of the open-source community.

To compare this GitHub data with, I have decided on using data that describes a country's allowance for its citizens to participate in democracy, along with actual participation numbers, to see if there is any correlation between more democratically aligned countries and their likelihood to contribute to OSS. This data will come from the Economist's Intelligence Unit (EIU) Democracy Index for the year 2022, as this data weights each country's democratic performance using multiple factors[2]. This includes their free and fair elections score (`elect_freefair_eiu`) and their democratic participation score (`pol_part_eiu`). This data sets scores each individual trait on a scale from 0.00 (least democratic) to 10.00 (most democratic).

While I could have used actual volunteering populations of countries, I decided this was not very representative towards the goal of my project. This is due in large part to how the number of volunteers a country sees yearly does not accurately describe their volunteering performance over time. For example, some people may volunteer more than once a year and be doubly counted. Some countries may also not have very recent volunteering data, an issue I ran into frequently while seeking the data out. Another issue I ran into was inconsistency between each country as to what actually constituted as "volunteer work", for cultural reasons or otherwise.

[1] J. Wachs, M. Nitecki, W. Schueller, A. Polleres. "The Geography of Open Source Software: Evidence from GitHub." Technological Forecasting and Social Change (2022): https://www.sciencedirect.com/science/article/pii/S0040162522000105

[2] https://github.com/owid/notebooks/blob/main/BastianHerre/democracy/datasets /cleaned/eiu_cleaned.csv

Democratic performance data, on the other hand, is much more heavily documented. There were many Democracy Indexes I could have drawn my data from; the only reason I chose the EIU's is because it distinctly separates electoral participation and democratic freedom, which I will use when weighting my results.

3 Hypothesis Testing

I intend on weighting the open-source contribution numbers of each country against their overall population numbers in order to describe their populations' participation in open-source projects. Then, I will compare this aggregate data against three different metrics.

The first metric will be the overall political participation index, calculated by the EIU and included in the dataset under the column `pol_part_eiu`. This describes the actual participation by citizens in elections.

The second metric will be the democratic culture index. That is, the extent to which a country's citizens prefer democratic institutions over non-democratic ones. This is found under the column `dem_culture_eiu`.

The third and final metric will be the free and fair election index, found in column `elect_freefair_eiu`. This describes the extent to which a country hosts free and fair elections for its citizens.

In order to properly compare this data, I aim to create multiple 2D scatterplots in order to possibly uncover a linear correlation between democratic participation and open-source participation. While not terribly complicated, I still do believe it will be an interesting thought experiment.