CS 5306 - Project 1

October 4, 2018

Hannah Lee (hl838), Hyun Kyo Jung (hj283), Niranjan Ravi (nr339)

Background

Our target crowdsourcing system is GitHub; there are many open-source projects publicly available on GitHub that are maintained, modified and updated by various developers. Among the millions of open-source projects on GitHub, some of them have become extremely popular to the general community, such as NumPy, React, and Bitcoin, to name a few.

We are interested in how this popularity affects the behavior of the contributors - especially the ones who had been contributing to the project even before it became popular. For example, would early members of the project stop caring about the project once the project gains popularity, because perhaps the project is not cool anymore now that it has become too mainstream? Or would they contribute even more now they feel more responsible for the project? Or it might be even the case that popularity doesn't in fact have any effect on the behavior of early contributors.

Our question

Our question is therefore: Do early contributors tend to contribute more or less once a project has become popular?

Data

We look at Gitential Datasets for Open Source Projects that is readily available online.

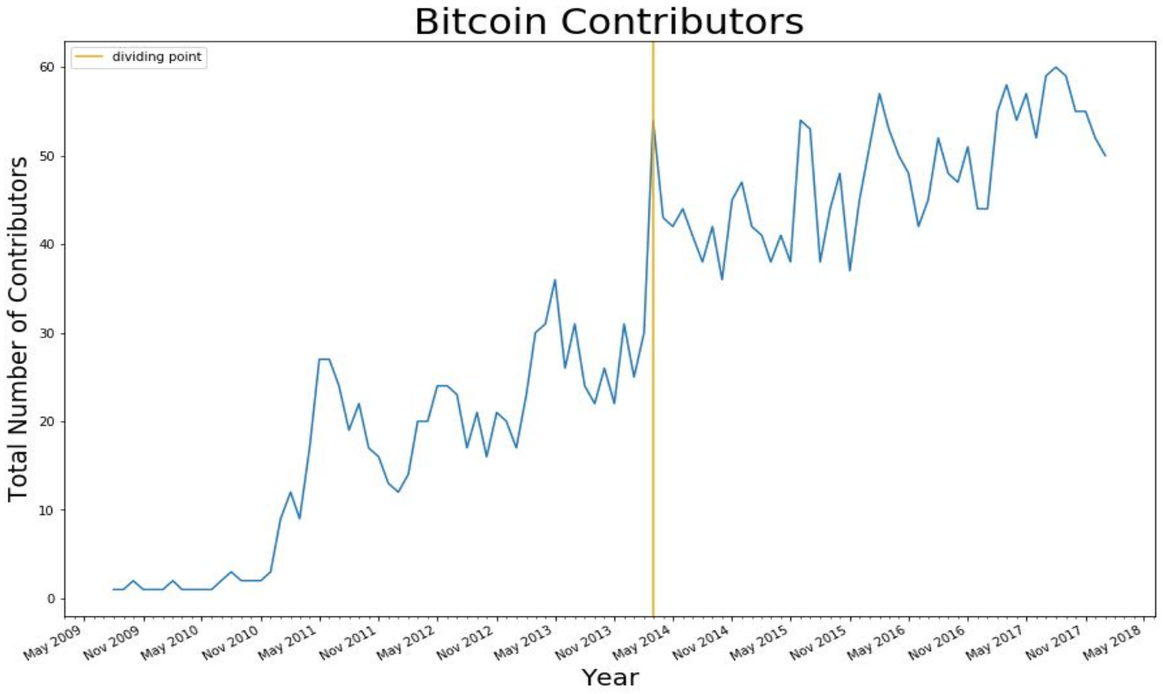
- We use the "commit data", which is a json file containing individual contributors' commits along with some metadata like time, author, lines of codes modified, etc., for different open-source projects.
- We measure an individual's *contributions* in terms of the number of commits.
- We compare early contributor's behavior on three different open-source projects - NumPy, Bitcoin, and React, each of which have been around for different amounts of time: 17 years, 9 years, and 5 years, respectively.
- The table below shows the total number of commits, total number of contributors, dates of the very first commit and the last commit for Bitcoin, NumPy, and React projects. (Note: last commit refers to the last time this data was extracted, not the last commit of the project).

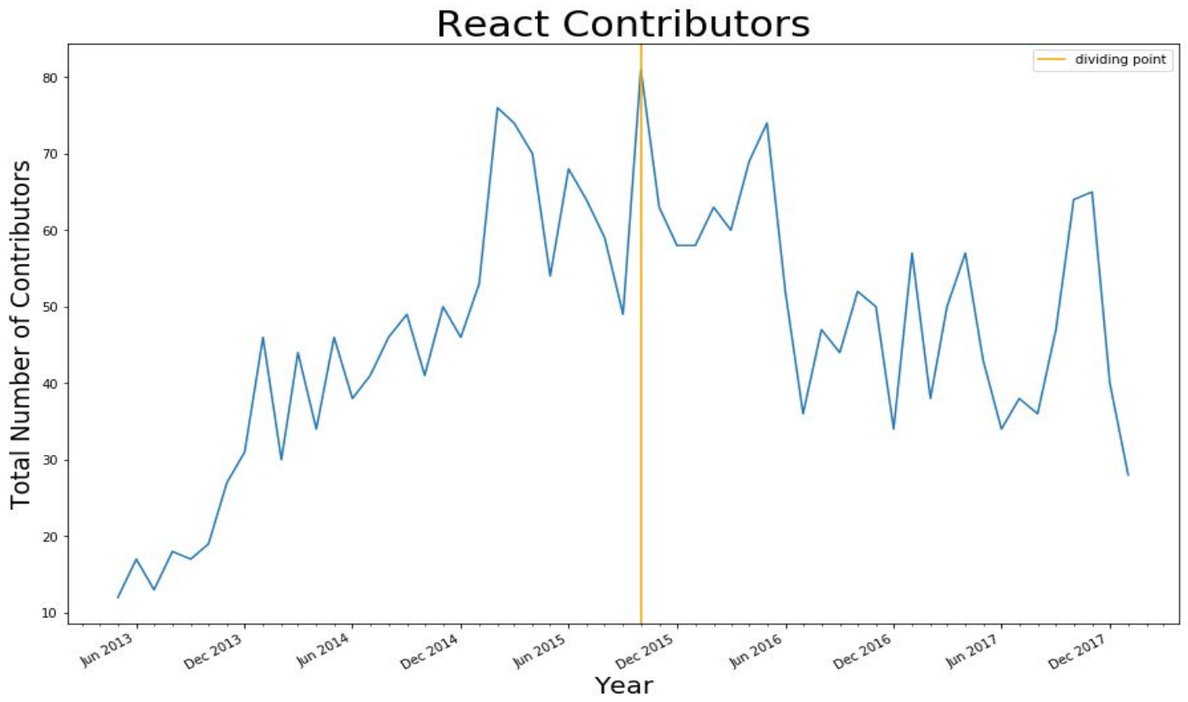
	number of commits	number of contributors	first date	last date
bitcoin	36302	812	2009-08-29 23:46:39	2018-01-18 10:05:40
numpy	24552	751	2001-12-18 10:45:10	2018-01-18 10:05:40
react	24157	1358	2013-05-29 08:54:02	2018-01-18 10:05:40

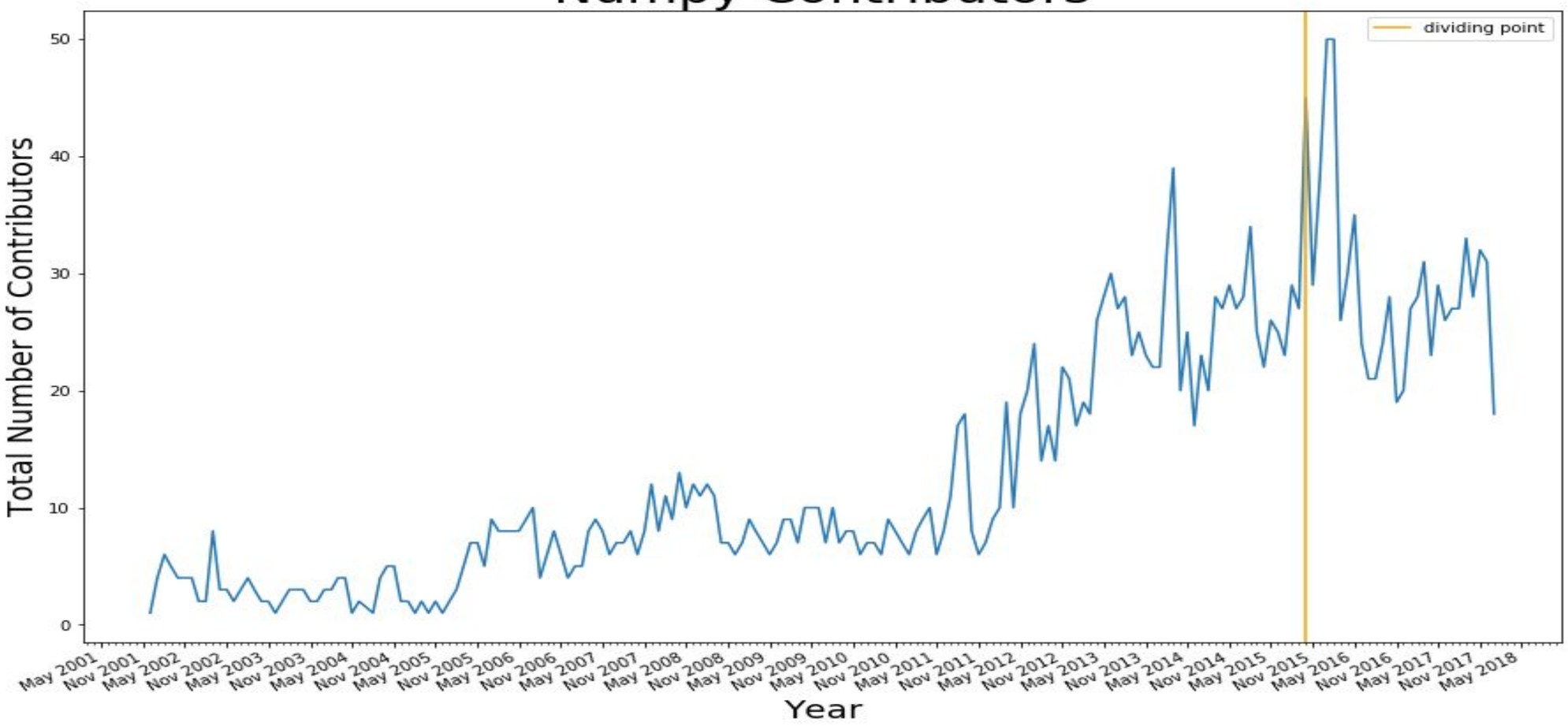
Visualization Results

In order to answer our question, we address the following:

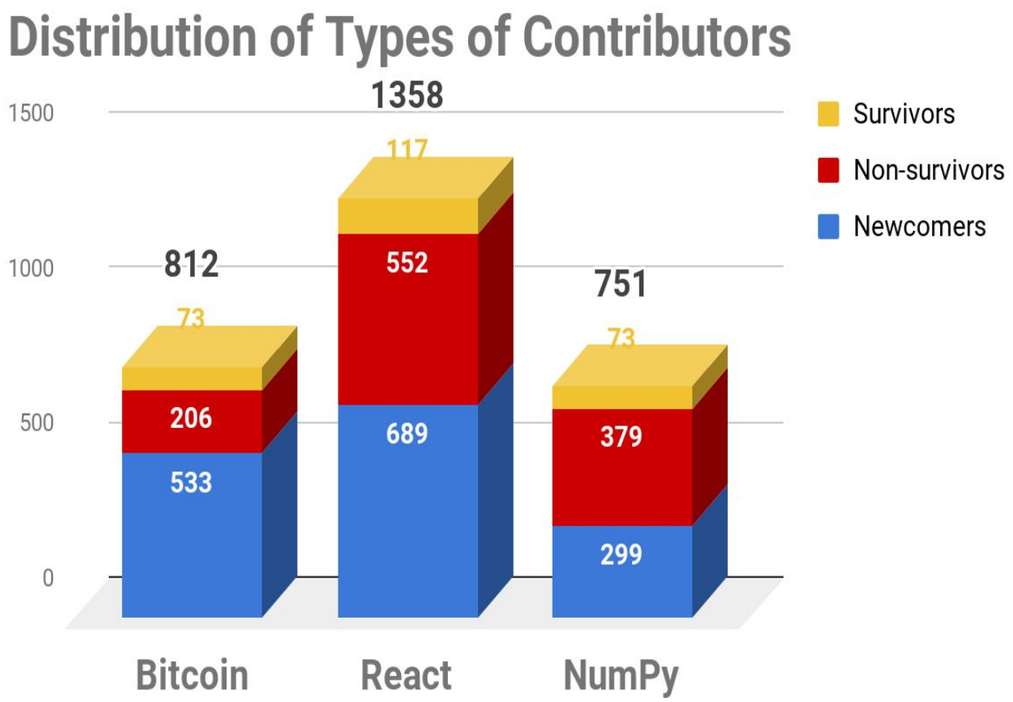
1. First, we define the term "early contributors" more specifically. We hypothesized that the project's popularity peaks in the month when there is the largest increase in the number of participating contributors. We call this month the ***dividing point,*** and use it to distinguish between the ***early contributors*** - people who started to contribute anytime *before* the dividing point - and ***newcomers*** - people who started to contribute anytime *at the time of* or *after* the dividing point. The below graphs show the total number of contributors for each month.

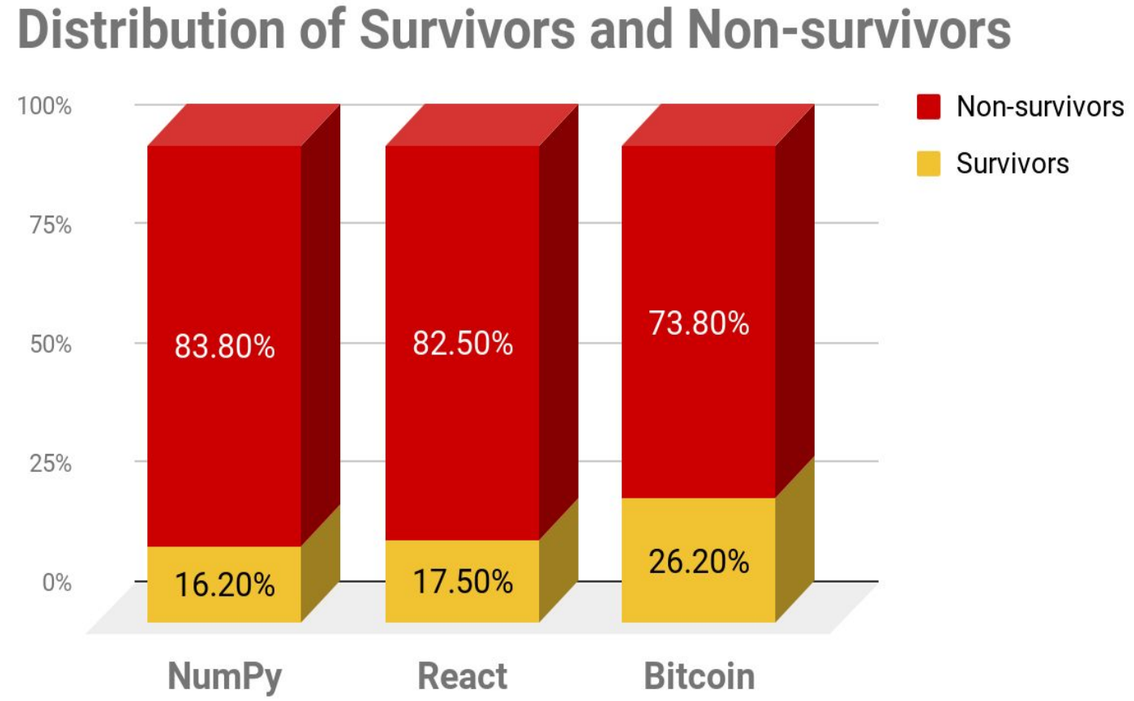

Bitcoin Contributors
dividing point
Total Number of Contributors
Year
May 2009
Nov 2009
May 2010
Nov 2010
May 2011
Nov 2011
May 2012
Nov 2012
May 2013
Nov 2013
May 2014
Nov 2014
May 2015
Nov 2015
May 2016
Nov 2016
May 2017
Nov 2017
May 2018


React Contributors
dividing point
Total Number of Contributors
Year
Jun 2013
Dec 2013
Jun 2014
Dec 2014
Jun 2015
Dec 2015
Jun 2016
Dec 2016
Jun 2017
Dec 2017

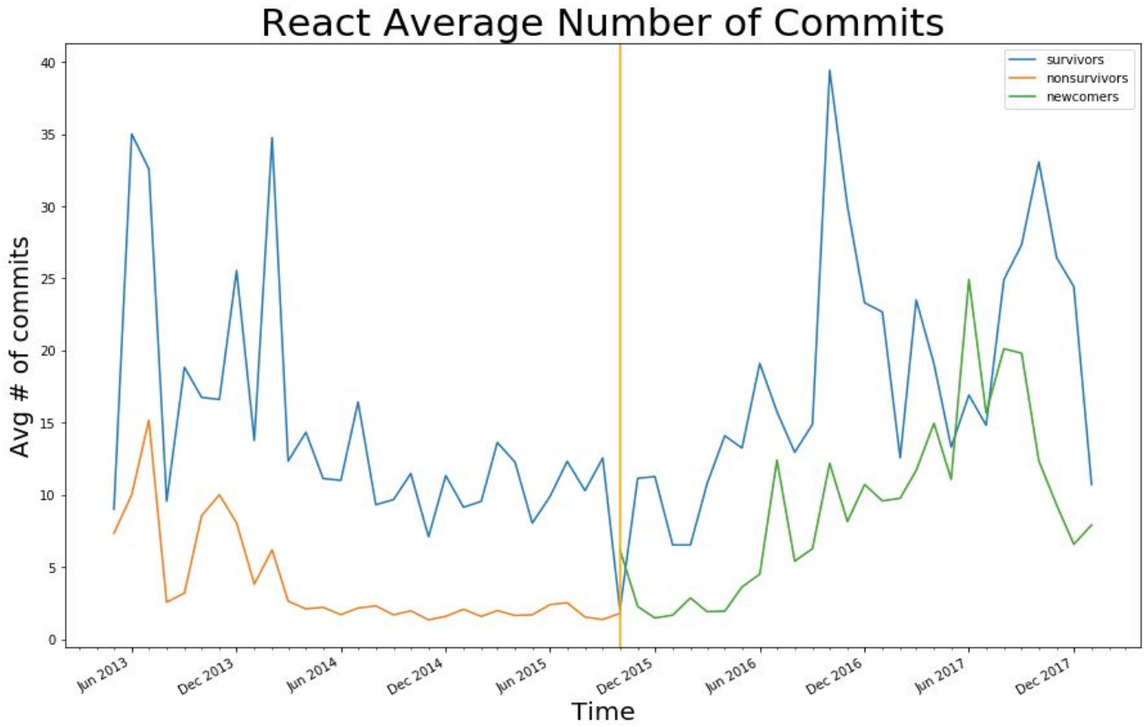

Numpy Contributors
dividing point
Total Number of Contributors
Year

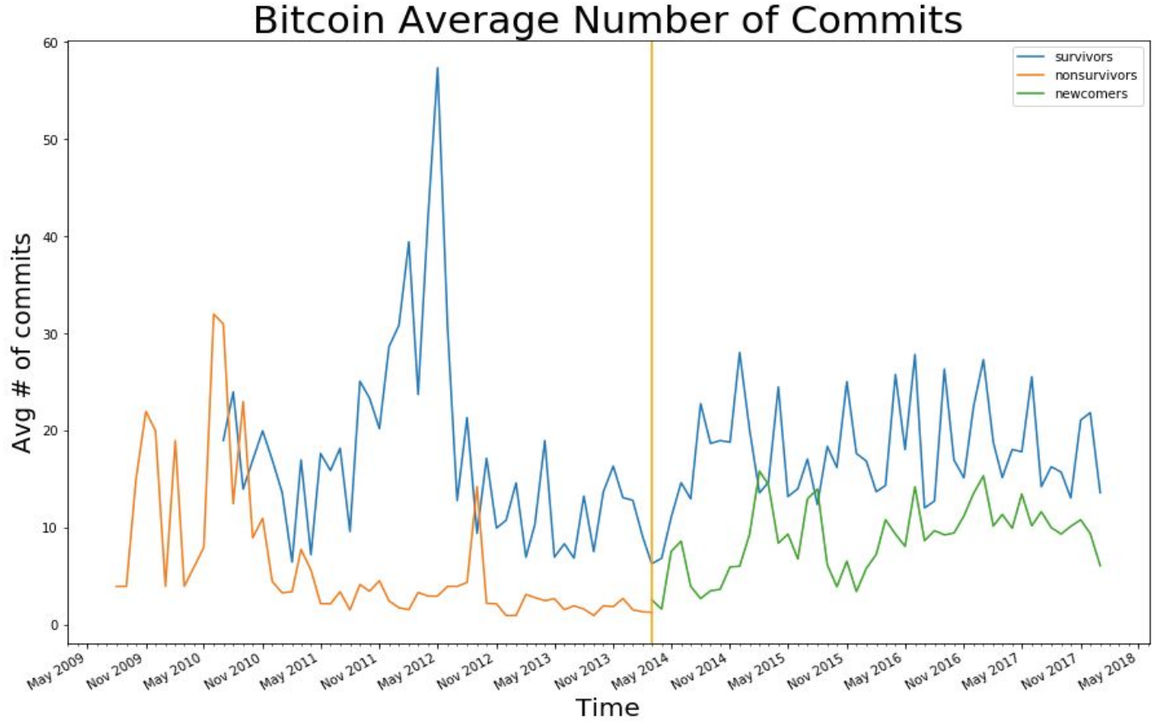
2. Among the early contributors, some left even before the project became popular while others remained in the project after the dividing point. To capture the early contributors' behavior more concretely for both of these subgroups, we further categorized early contributors into ***survivors*** and ***non-survivors***: *survivors* are those who continued to contribute *after* the dividing point, and *nonsurvivors* are those who stopped contributing *before* the dividing point.
    - Left: raw count distribution of early contributors, specified into survivors and non-survivors, and newcomers for each project.
    - Right: distribution of survivors and non-survivors among the early contributors for each project. We see that there are a lot more non-survivors (~80%) than survivors.

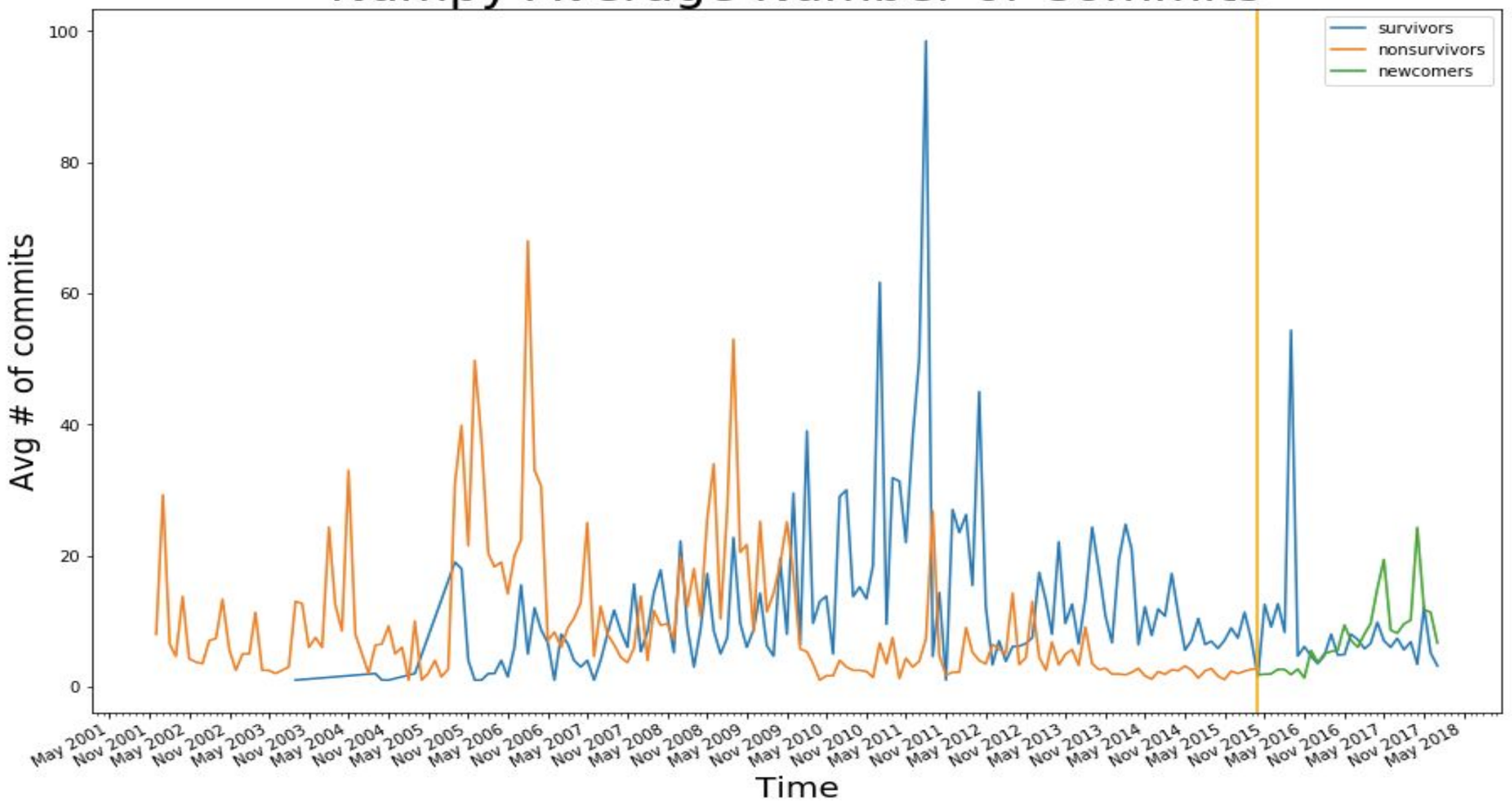




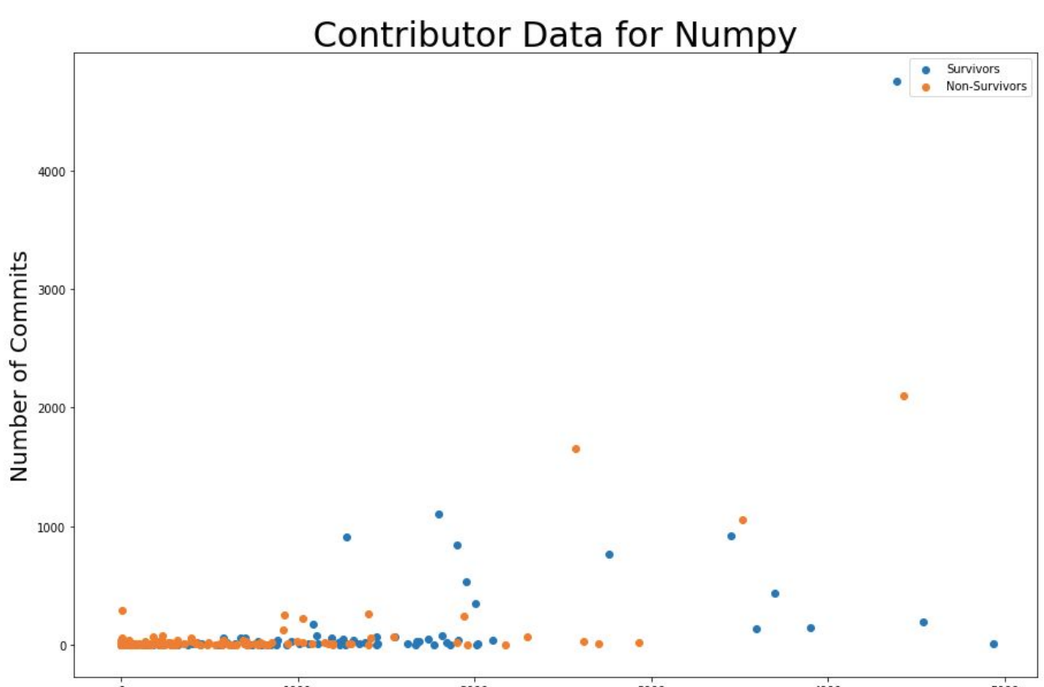
3. Now, we're ready to answer our original question: "do early contributors tend to contribute more or less once a project has become popular?" We look more closely into the contributions made by survivors and non-survivors separately.
   The below graphs show the average number of commits per month for each of the subgroups: survivors, non-survivors and newcomers.
    - **Survivors** (blue line)
        - Before the dividing point (survivors vs. non-survivors)
            - survivors generally contribute a lot more to the project than the non-survivors, except for NumPy where the difference of contributions between the survivors and the non-survivors is not very clear.
        - After the dividing point (survivors vs. newcomers)
            - We see that the survivors in React and Bitcoin consistently contribute a lot more than the newcomers, while in NumPy, the survivors' contribution increases dramatically right after the dividing point and then decreases below the newcomers' contribution level.
    - **Non-survivors** (orange line)
        - In general, the non-survivors seem to be interested in the project only during the early phase of it, as demonstrated by the spike in the graphs. However, their average number of commits is generally upper bounded by the average number of commits by the survivors for each month.
        - So, we conclude that the non-survivors are the ones who are excited about the project in the beginning but gradually lose their interest as time goes on, and eventually leave even before the project peaks in popularity.

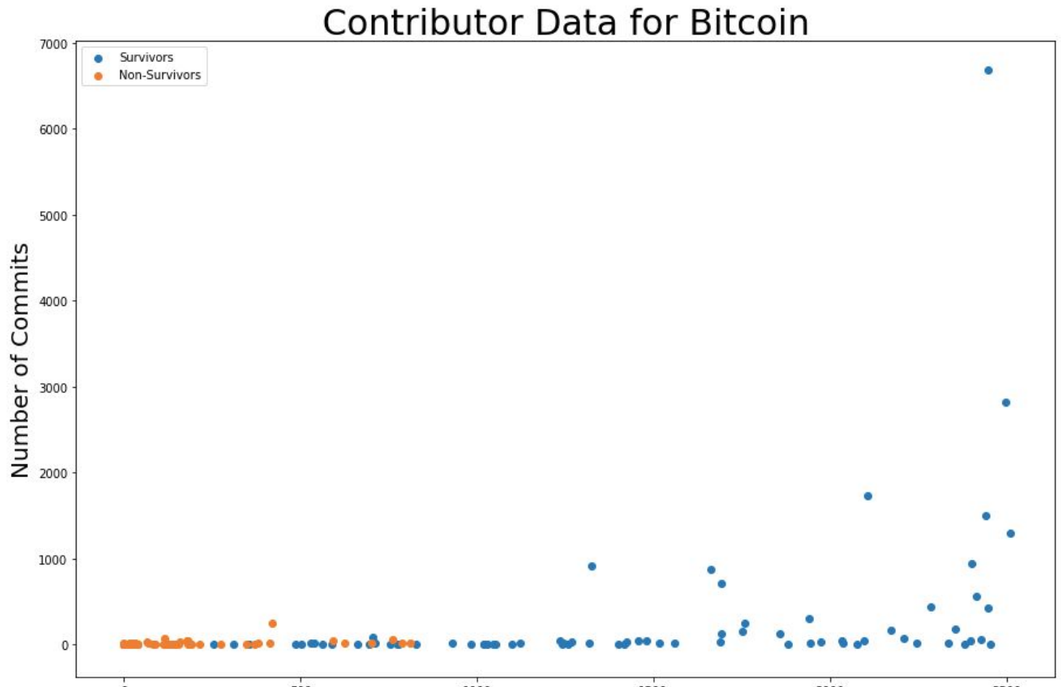

React Average Number of Commits
survivors
nonsurvivors
newcomers
Avg # of commits
Time
Jun 2013
Dec 2013
Jun 2014
Dec 2014
Jun 2015
Dec 2015
Jun 2016
Dec 2016
Jun 2017
Dec 2017

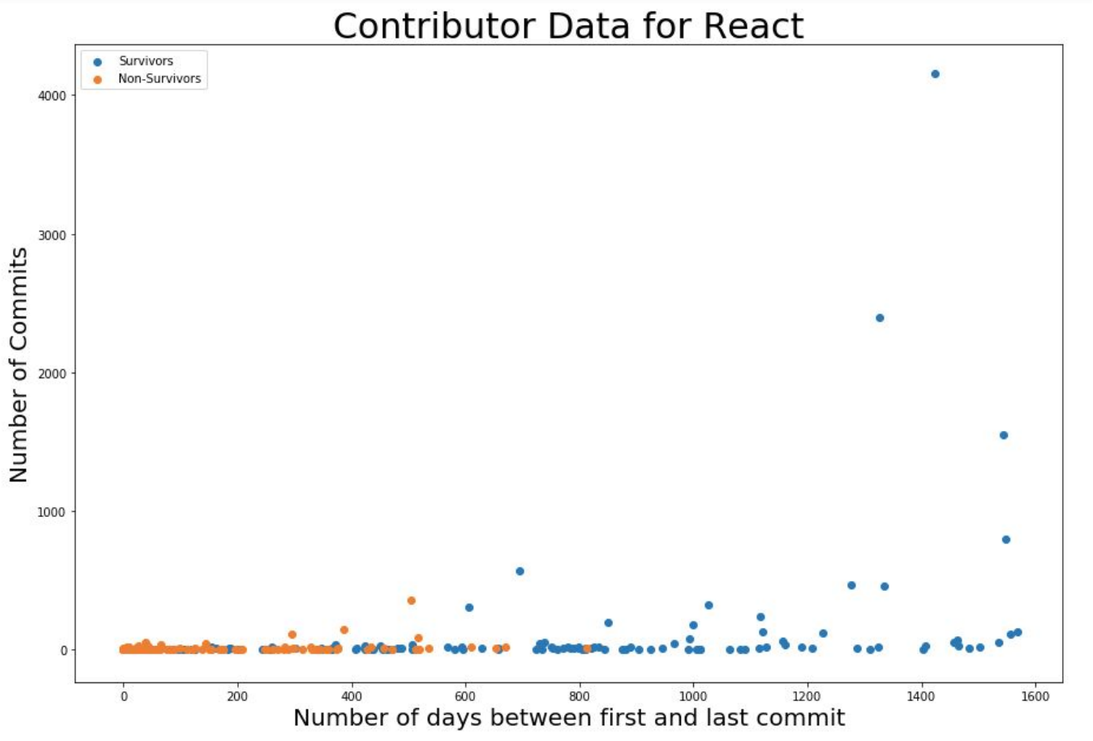

Bitcoin Average Number of Commits
survivors
nonsurvivors
newcomers
Avg # of commits
Time
May 2009
Nov 2009
May 2010
Nov 2010
May 2011
Nov 2011
May 2012
Nov 2012
May 2013
Nov 2013
May 2014
Nov 2014
May 2015
Nov 2015
May 2016
Nov 2016
May 2017
Nov 2017
May 2018


Numpy Average Number of Commits
survivors
nonsurvivors
newcomers
Avg # of commits
Time

4. Additionally, we wanted to see if most of the survivors are simply the ones who joined the project right before the dividing point, and the non-survivors are the ones who started contributing early in the beginning who just got tired of the project after a while. So, we investigate how long each survivor and non-survivor stayed in the project (as measured by the time difference between their very first commit and the last commit), as well as how often they committed to the project.
    - We find that survivors all remained active for longer than the non-survivors, with a few exceptions in the case of Numpy, since its dividing point was 14 years after it began.







Summary

Ultimately, the data is not conclusive enough to show a general pattern for the early contributors' behaviours after the projects become popular, as it varies from project to project. However, the data does seem to show that in general, the survivors made more contributions to the project than the non-survivors and newcomers. Once again, the data for Numpy is a bit skewed, since we have 14 years of “pre-popularity” data, but only 3 years of “post-popularity” data. It will be interesting to see how these results hold up in the long term, as the newcomers begin to outnumber the survivors.