the WINNOWER

SCIENCE AND SOCIETY

Correlating the Sci-Hub data with World Bank Indicators and Identifying Academic Use

BASTIAN GRESHAKE

CORRESPONDENCE:
bgreshake@googlemail.com

DATE RECEIVED:
May 30, 2016

DOI:
10.15200/winn.146485.57797

ARCHIVED:
June 02, 2016

KEYWORDS:
sci-hub, copyright, academic use

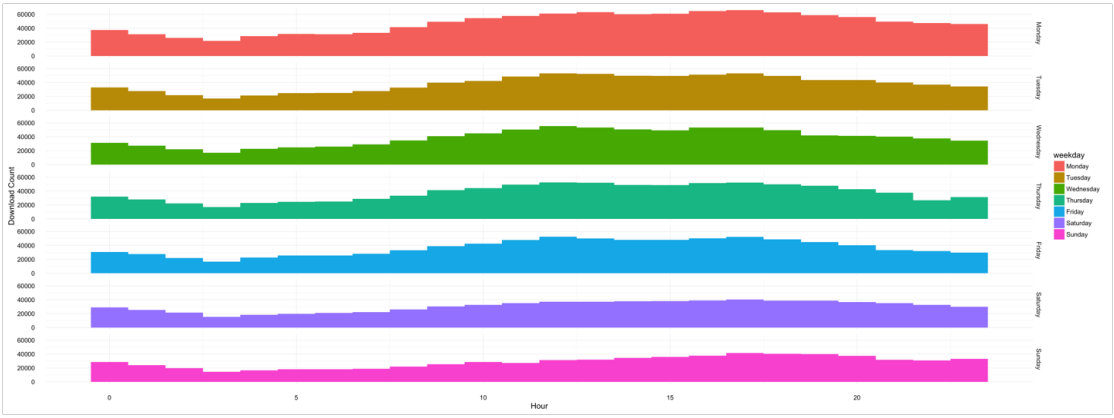
CITATION:
Bastian Greshake, Correlating the Sci-Hub data with World Bank Indicators and Identifying Academic Use , *The Winnower* 3:e146485.57797 , 2016 , DOI: 10.15200/winn.146485.57797

This was originally published as two separate blog posts at my blog:part 1 & part 2.

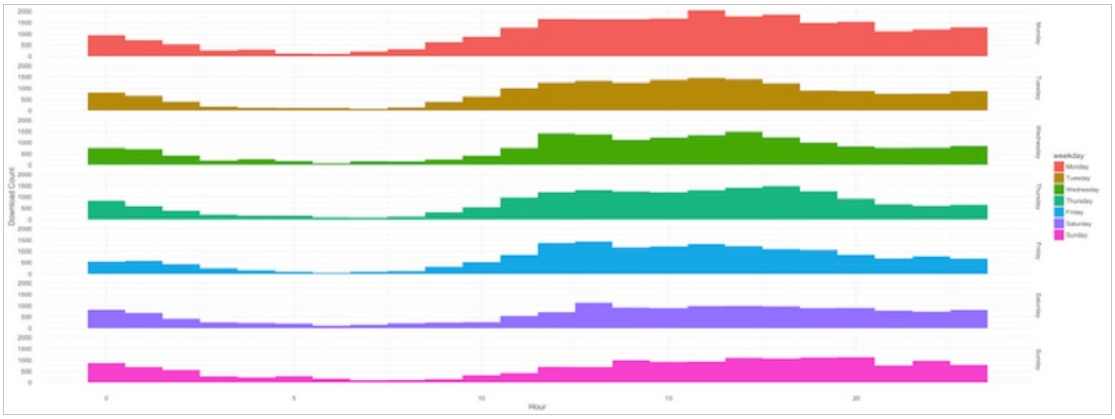
If you're following the world of academic publishing (and the paywalls that keep researchers from doing their jobs) you will have heard of Sci-Hub, which offers an elegant way to get around those paywalls. Often called *the PirateBay of Academia*, Sci-Hub gives you access to over 48 million paywalled articles, happily ignoring copyright. In my book that's a great thing to do, after all it's the general public that's paying academic research and they are the ones that suffer most from paid access. And don't get me started about the absurdity of university libraries having to buy back the results their own institutions generated.

In any case, between September 2015 and February 2016 over 28 million downloads were done using their system. And at the end of April 2016 Sci-Hub published the meta-data for those downloads, along with analyses done on the data and very cool visualizations. The data includes not only the time the download was done, but also the DOI that was requested (which also tells you which journal/publisher was paywalled) and the country/city from which the download was made.

When I saw this data I immediately wondered whether most of those downloads were made from outside universities/by the general public or whether they were done by academic researchers who couldn't get access through their institution due to the lack of subscriptions. So I started digging into the download times, looking at which weekdays and times most downloads occurred.



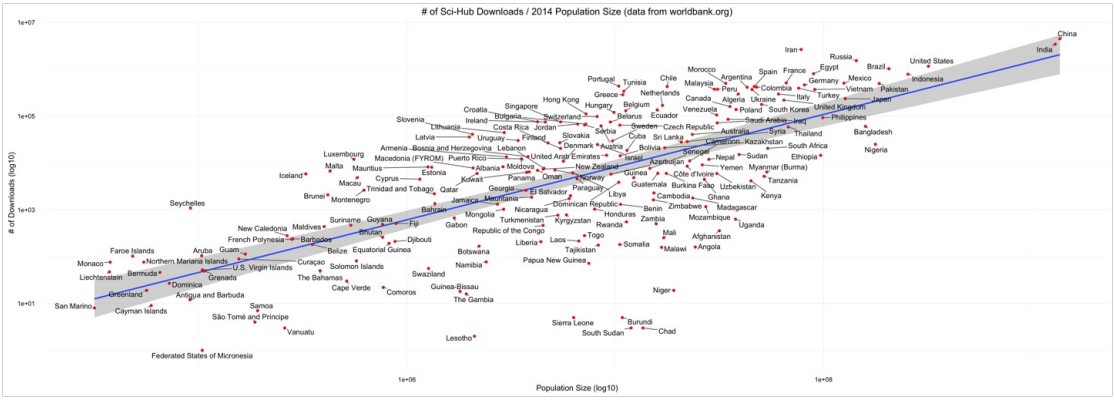
The bulk of downloads seems to take place between the magical 9am to 5pm time frame, which means people are accessing the data during their work hours, which will include many academics. So it's not only people after hours (either the general public or academics from home) who access the data.



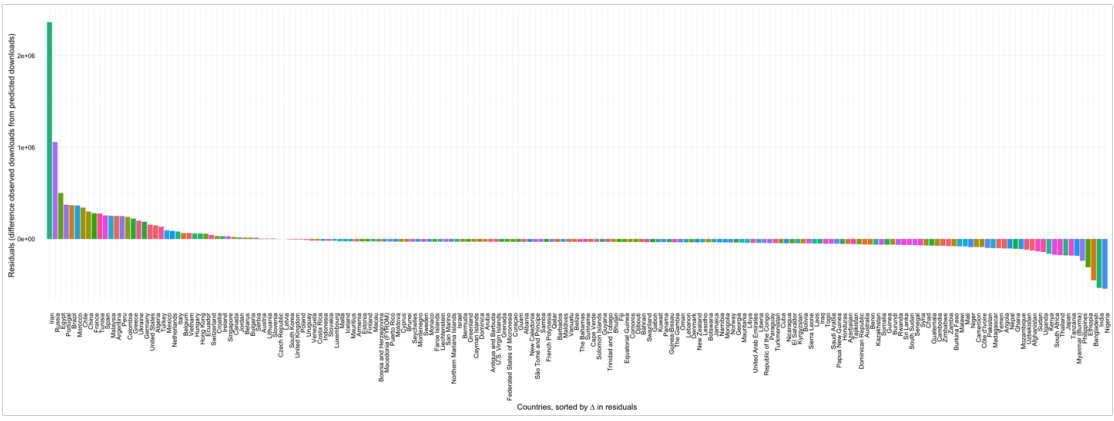
A small aside: As rightfully pointed out on Twitter, I used all data for this, but the time stamps are all UTC, so non-UTC time zones distort the picture. To account for this effect in the quickest way I plotted the same data, filtered once for downloads from Germany (above) and once for downloads from Hong Kong (below). You see the shift in those distributions, but if you adjust for it the picture stays the same.

But the country data can be used for so much more things than just adjusting the time zones. You can easily correlate the data with more general public data sets. For example, the World Bank offers a great deal of public data. You can get the population for all countries, the GDP or the number of internet users.

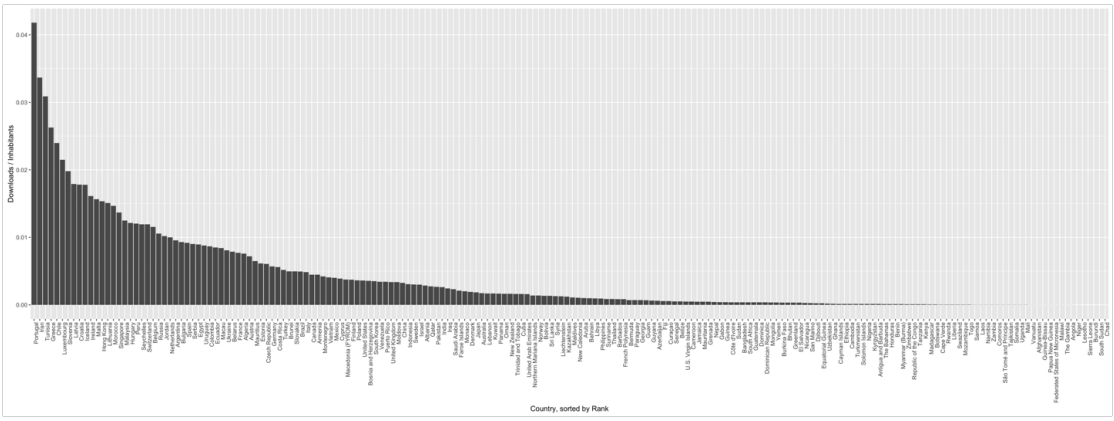
I started out by correlating the population sizes to the number of downloads. You'd naively expect that the larger a population is, the more downloads will be done. Which is true, up to a point. If we plot the population size and the number of downloads (on a log10 scale), we see a nice positive correlation between those.



But there are obvious outliers. Some countries, such as the Iran, have much more downloads than expected given the linear fit, others have much less. We can look at the residuals if we want to see how much those countries differ from the expectation given by our model. This is just the difference between the expected value and the observed value.

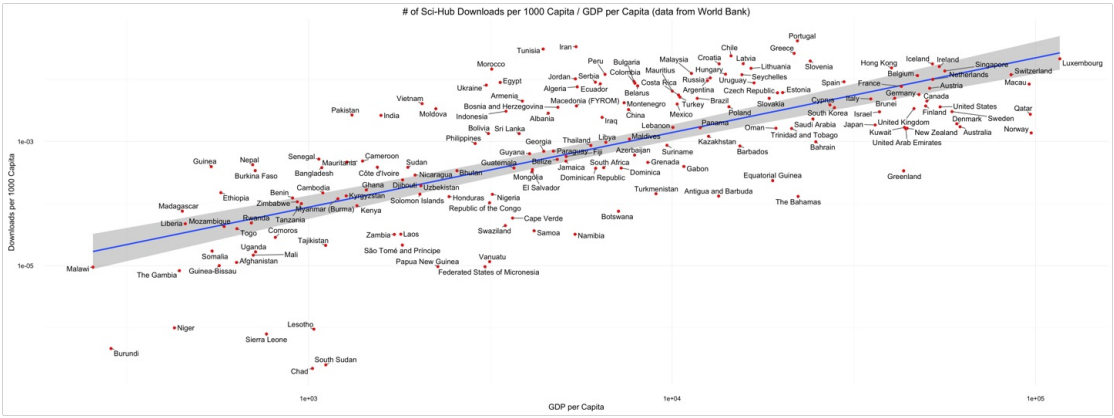


We again see, when looking at the absolute difference, that much more downloads are done from Iran than we'd expect. And Scott pointed out that some of the Iranian mirrors were not even included. When we move away from those total numbers and instead look at *Downloads per 1,000 inhabitants* we see a pretty similar picture.

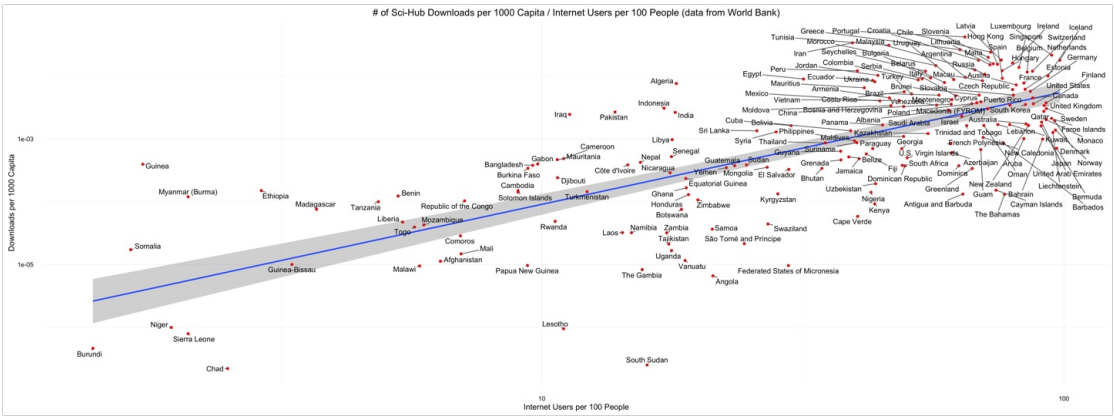


Using the *Downloads per 1,000 inhabitants* metric we can now also correlate those numbers to other

metrics measured by the World Bank. For example the Gross Domestic Product (GDP) per inhabitant. Which yields a nice positive correlation as well, basically saying that the higher the economic performance of a country, the more people will illegally download scientific publications.

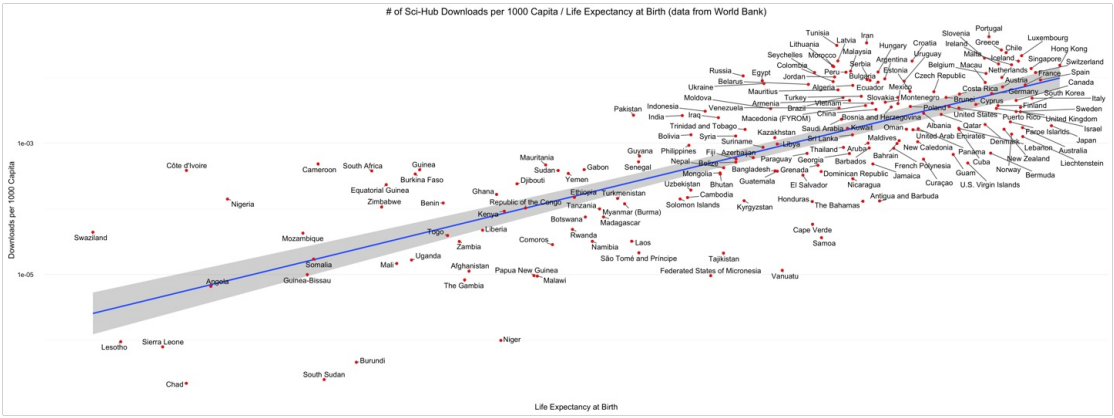


But is it really the GDP in itself that's driving the effect, or are we rather measuring other things that might explain it? What about being online? In order to illegally download publications you need to be able to download stuff at all.

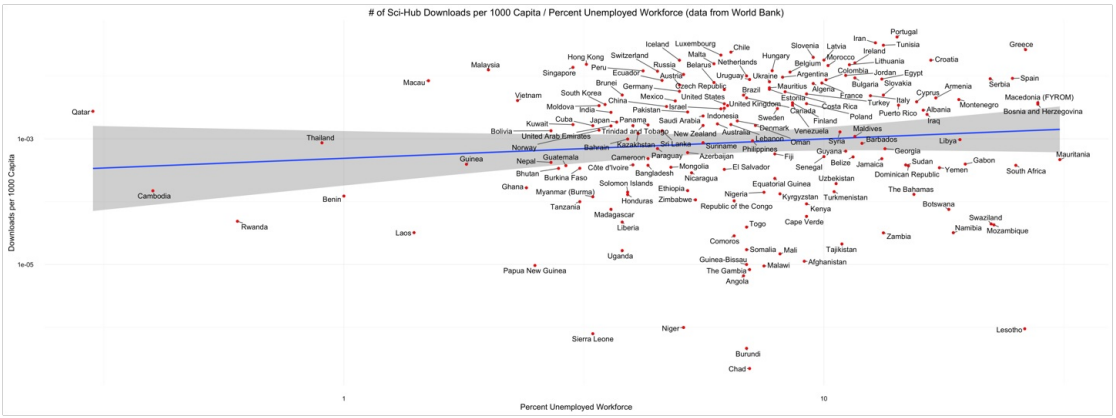


So far this is pretty much what you'd expect: The more people are online, the more downloads you find for a given country. Which also isn't too surprising, if you think about that GDP and the number of internet users itself is closely correlated to each other.

Another thing that obviously could have an impact on whether you'd be interested in academic literature is whether you have time to care about such things. So let's look at life expectancy at birth.



This also seems to intuitively make sense. The higher the life expectancy, the more downloads you find for a given country. Let's look at one last metric here: The unemployment rate amongst the potential work force.



Which, interestingly enough, is a very bad predictor of how many papers will be downloaded. Regardless of unemployment levels there are countries which do download a lot of papers and others less so. Amongst the high-unemployment countries Spain and Greece make interesting positive outliers for example, having lots of downloaded papers.

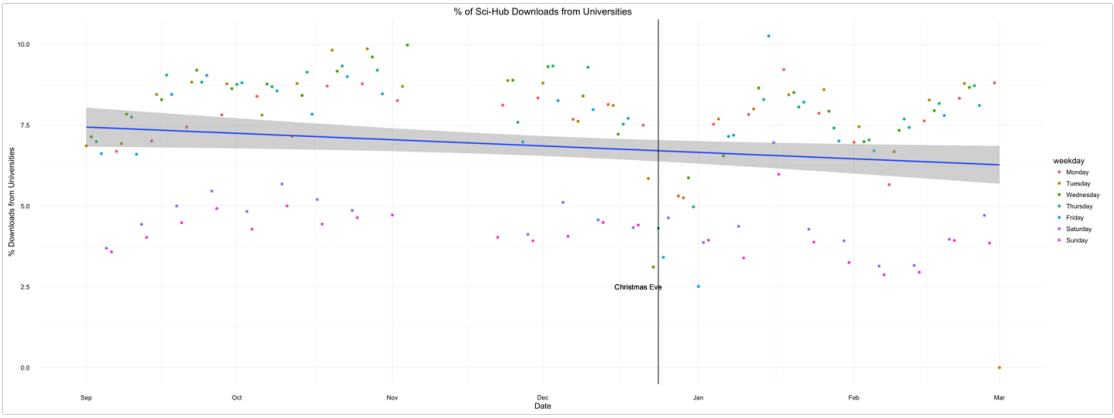
Unfortunately this data so far does not tell us much about how Academics are using *Sci-Hub*. Or more specifically: **How many people are actually using *Sci-Hub* to download publications while they are in universities?** Because one of the claims that I (at least used to) hear quite often when debating open access is that academic researchers have nothing to complain about, because they can access all the literature in any case thanks to library subscriptions.

While this claim obviously ignores all the legitimate reasons for non-researchers to read primary literature (like, patients reading up on their illnesses – or just the general public who paid for all the research to start with…), I wondered: Is it even true? From my own experience I was pretty certain that it's not the case, and I guess it's not only because the University of Frankfurt is too cheap to pay for subscriptions.

A nice approach to my question would be to cross-reference the *Sci-Hub* data with known university IP-ranges, to see which downloads are actually made from inside universities. But for straight-forward reasons the *Sci-Hub* data didn't include any IP addresses. So I did the same thing that John Bohannon did for his piece in Science: I asked for the data.

Because if you look around a bit (and nicely ask some people to help you in your search) you can find a list of around 5800 University/College IP ranges from many places around the globe. This list obviously has some shortcomings: It's about 1 1/2 years old by now and the names of the universities are mined from *LinkedIn* and *Webometric*. So I would say it's pretty safe to say that we are missing out lots of universities and their respective IPs with this list, so all measurements generated from this list should be taken as a lower bound.

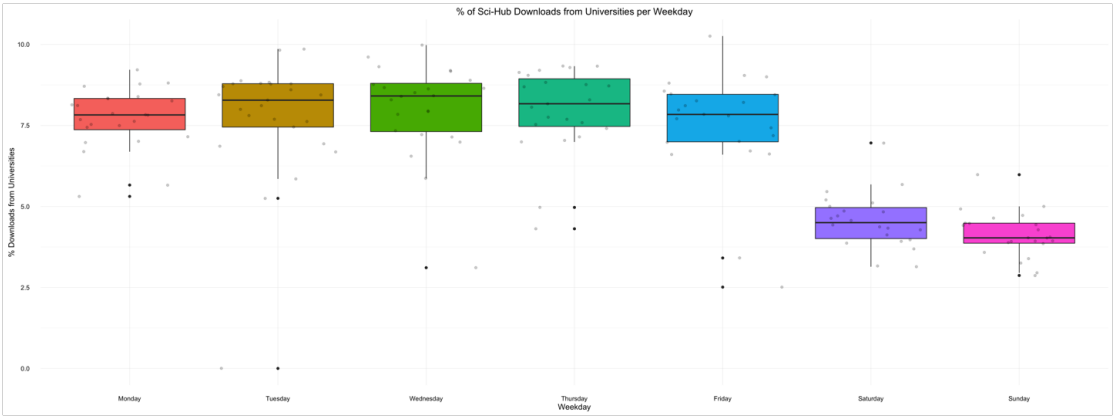
In any case: These public address ranges could then be matched to the data held by *Sci-Hub*. They were so helpful to send me the anonymized and matched data in two levels of detail: The first data set contains the percentage of downloads done from those university addresses for each day. With this it is easy enough to see that academic use of *Sci-Hub*, while not completely stable, seems to oscillate between 8 and 10 percent on a given workday.



As you would expect (or at least hope), the number of downloads from universities does cut back over the weekends (and holidays like Christmas). Jonathan Ready, who read my initial blogposts, also emailed me with an idea for why one can see seasonality in the percentage of academic downloads:

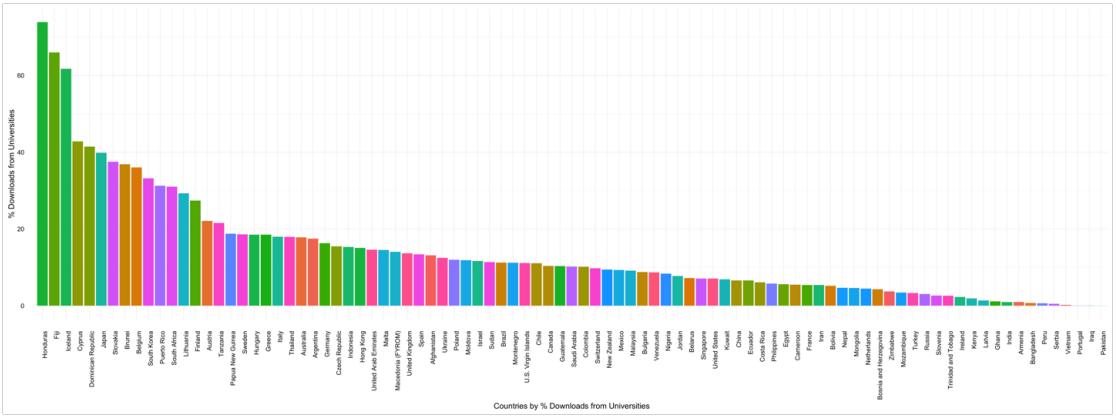
> *You may already have many replies, but I guess the seasonality is due to the influx and egress of post-graduates during the year - dissertations tend to be written at certain times of the year and so there is a great rush on citation searches at that time. This clearly varies between countries but I am sure it is likely a major cause of this variation.*

One important thing to keep in mind: This data is the worldwide aggregate, so different time zones will play into the weekday-calculation (and not all cultures hold their weekends on Saturday/Sunday).

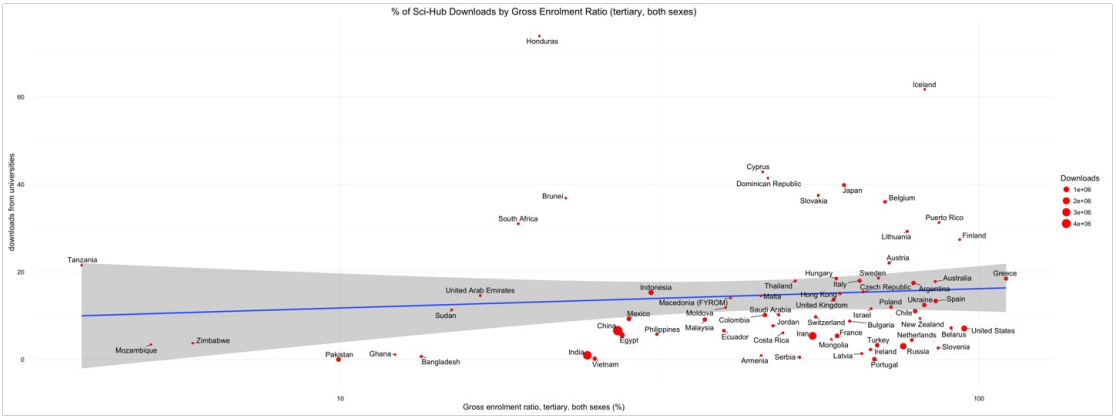


If you directly compare the different weekdays to each other, you can see the weekend effect even better. All in all we might draw two conclusions from this: Firstly, academia never really sleeps. And secondly, **we can answer John Bohannon's question on *Who's downloading pirated papers?* with a resounding *Academics do for sure!*.**

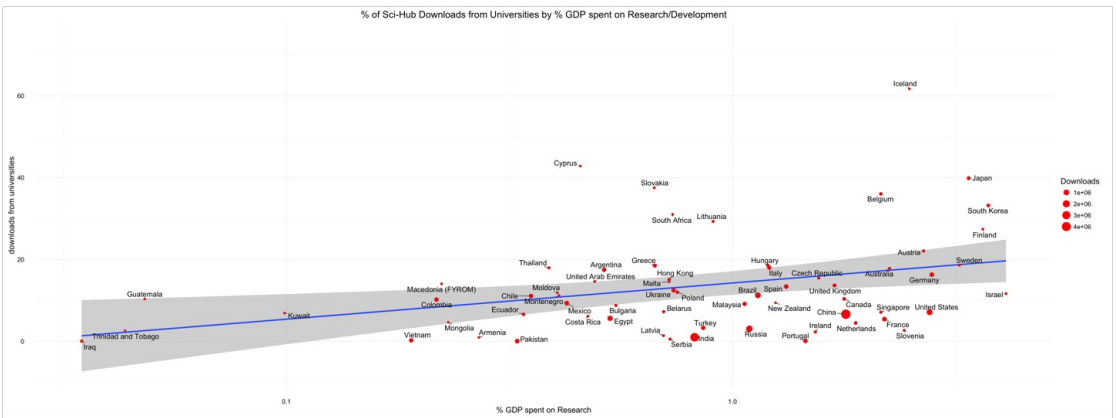
The second data set that I received on the academic use of *Sci-Hub* goes down to the country level. But to protect the people using the service, the resolution of the data has been cut back. So instead of providing 24h resolution, data is grouped by 10-day intervals. But this still allows to compare the academic use between different countries, at least to some degree.



If we just plot the mean academic use for the countries included in the data we see a huge spread, with academic use ranging from 70% to basically 0% (the plot excludes countries with 0% academic use). The huge variation is in part due to the fact that some countries have very small sample sizes. Front-runner Honduras for example only has 1,070 downloads in *Sci-Hub* in total, so a hand full of academic users can skew the value. On the other extreme, Peru for example does have a lot of downloads in principle, but the IP range data only includes 5 entries. Which might very well be an underestimation of Peru's higher education landscape and thus be extremely misleading.



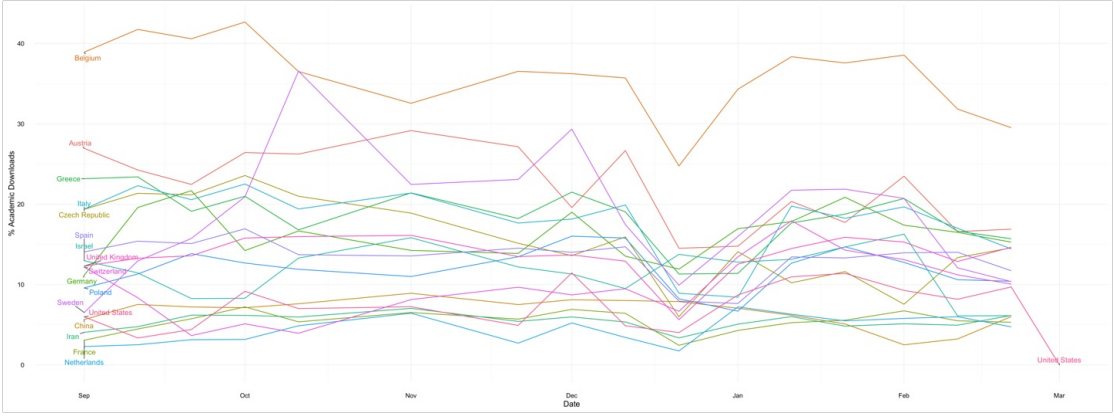
I nevertheless tried to see whether there's some underlying reason for this distribution. The first idea I could come up with was the percentage of the population being enrolled in higher education. Again the Worldbank was useful to find some data on this. I used the *Gross enrolment ratio for tertiary education across both sexes* (title by the Worldbank, not mine). Plotting this against the percentage of downloads (and at least visualizing the bias in the data by plotting the total number of downloads alongside), we see very little influence of it. Which might be because the gross enrolment includes non-university level education, but I somehow doubt that this has much of an influence.



The second hypothesis I had was that it's dependent on the amount of spending done on research in general: Once you start investing into research, having more academics around might lead to an initial increase in *Sci-Hub* downloads, which eventually might decrease because once you have enough money, you can afford the journal subscriptions? But that also doesn't seem to be the case. I guess at the end it does not matter how much money you throw at publishers, you'll never be able to pay their subscriptions.

The data used for these plots, as well as the code for generating the plots is on GitHub, as usual. If you have ideas on how to improve the analyses drop me an issue there or tweet me. And big thanks go to Alexandra Elbakyan for providing the data, I hope seeing this is of some use.

Post Script After tweeting a bit with @MaliciaRogue about my initial blog posts I wanted to see how the download behavior in different countries changes over time. I not-so-randomly selected a couple of European countries that came to my mind and also some potentially interesting other countries for a small plot.



There are some cultural differences to see here, or at least that's what I'd assume. For countries that have been heavily influenced by Christianity you can clearly see the Christmas holidays again. And looking at the weeks before Christmas in the Netherlands, the United States, Sweden and also Germany I somehow get the feeling that the protestant work ethic might be to blame for those pre-Christmas peaks.

Not too surprisingly Israel and China's numbers are pretty stable over Christmas. But Rosh Hashanah, Yom Kippur and Sukkot can be clearly seen leaving their marks, as can Chinese New Year. Are there any more holidays you can spot?