Longitudinal COVID-19 analysis of DNS traffic

Niklas Karlsson Joakim Thorn Maximilian Vorbrodt Hampus Rosenquist
Linköping University
Sweden
{nikka560,joath015,maxvo113,hamro777}@student.liu.se

Abstract

Due to the COVID-19 pandemic people have been isolated in their homes and have therefore had more time to spend on the internet. However, it is still mostly unclear how this time has been distributed between different websites and categories of domains on the internet.

The aim of this study was to do a longitudinal analysis of how the popularity of different domains have changed during the pandemic compared to before. Therefore data between year 2019 and 2022 was collected. The study was conducted through analyzing datasets of top domain rankings divided into categories, which are based on mainly DNS data.

We found that the popularity of domains containing entertainment of various kinds, such as "News and Media", "Computer, Electronics and Technology" and "Adult", saw a sharp increase at the start of the lockdown, but later started to decrease in popularity gradually. "Gambling" however increase steadily over the whole period.

Domains related to "Home and Garden", "Health", "Jobs and Career", and also not surprisingly "Sports" and "Traveling and Tourism", saw a decrease in popularity during the entire COVID-19 pandemic, but are starting to increase in popularity once again as the pandemic decreases.

1 Introduction

The COVID-19 pandemic caused a global lockdown in large parts of the world, meaning people had more time to spend on the internet. What people spend their spare time on may have a large impact on people's well being and society at large. Providing data on changes of what websites people spend their time on during this period may therefore be of value for analysis.

In this study we analyze the data from two domain top lists, Cisco Umbrella and Tranco. Both contain a list of top one million visited domains however, while Umbrella is based on DNS data, Tranco is based on aggregation from several popular top domain lists.

In this paper, we show that although people initially turns to entertainment to fill out their time, they eventually seem to gravitate towards other activities. We also show that news outlets and social media platforms have steadily increased in popularity during the pandemic. The paper also contributes a starting point and methodology for conducting similar studies on the topic in the future, as it may be interesting to see how the popularity of domains will change when the pandemic has ended.

There are studies that have made use of DNS domain names and ranking lists before, as well as studies that have investigated the COVID-19 pandemic and its consequences. In this paper we combine the two and argue that this combination, should give an interesting insight into how social isolation affects the way people spend their time on the internet.

Furthermore, our analysis and data could prove useful to find user patterns across the internet to gain a better understanding of how isolation may affect human behaviour.

The paper is structured as follows:

- **Methodology** - Contains background information about methods, approaches and algorithms used to achieve the result.
- **Result** - Resulting graphs after having conducted the analysis of top lists.
- **Conclusion & Discussion** - Summarizes the key findings from the result and answers the question of how people have spent their time on the internet before and during COVID-19.

2 Methodology

In order to track both current and past trends, the use of domain top lists are a common tool for researchers. By comparing lists from before the pandemic broke out, during and after we can see how the list of most popular websites has changed and possibly spot changes in trends and patterns.

2.1 Cisco Umbrella

There are numerous domain top lists available to choose from. One such top list is Cisco Umbrella [5]. This list is based on DNS data gathered from all over the planet and ranks the most popular web domains based on the number of times they have been visited.

Conference acronym 'XX, June 03–05, 2022, Linköping, Sweden

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

As we wanted data that represented the web traffic both before the break out, as well as after, the time period of **2019-01-01** to **2022-03-31** was used in this study. Similar to Scheitle [2], we used the toplist database deployed at [2] to get access to the daily top lists needed. To fetch this data we wrote a python script which downloaded and unpacked every Cisco Umbrella list .csv file from the time period mentioned, as well as renamed them accordingly for easier processing later on.

2.2 Tranco

One problem with using top lists is the integrity of the data. Research has been made regarding biases within the top lists. Le Pochat et al. discuss the use of Tranco [1]. Tranco is another top list used by both researcher and service providers to analyze internet traffic and web trends. Tranco attempts to solve the issue of top lists being biased and manipulated by combining several top domain lists and calculating an average ranking for each domain [4].

This in turn results in a new, clean and arguably non-biased, or at least less-biased, list which can be used as a complement for researchers studying web traffic.

Tranco also comes with its own API which, similarly to the python script we wrote for Cisco Umbrella, could be used to download each daily list during our time period. Again, the .csv files were downloaded and renamed accordingly in order to be processed later. Worth noting is that the daily Tranco list for **2019-07-29** was unavailable and therefore not considered in this paper.

2.3 Categories

In order to track trends on the web, a number of categories containing some of the most popular domains within that category was setup. There are endless amount of categories that could be investigated, but to limit this to a reasonable amount the research team decided on 14 categories.

The categories decided on were

- Adult
- Arts and Entertainment
- Computers, Electronics and Technology
- E-commerce and Shopping
- Gambling
- Games
- Health
- Hobbies and Leisure
- Home and Garden
- Jobs and Career
- News and Media
- Science and Education
- Sports
- Travel and Tourism

The choice of categories were partially based on publicly available categories supplied by SimilarWeb [3], containing the 50 most popular domains within that category. The 50 domains of each category were manually scraped and added to their own text file.

The choosing of categories was also determined by the research teams opinion on the likelihood of the category being affected by the COVID-19 pandemic. That is for example, domains within the category "Heavy Industry and Engineering" is less likely to be affected than a category like "Arts & Entertainment". Although, it should be noted that all categories have connections to the COVID-19 pandemic, however it is outside the scope of this paper to investigate each and every category possible.

2.4 Dataset generation

Because we are looking at how trends are changing due to the pandemic, individual domains are of less interest and less importance than the chosen categories. To study these categories we created new daily .csv files in which we only look at the 700 domains gathered from the SimilarWeb lists. We also wanted to ensure that all domains are occuring in both of our source files (Cisco Umbrella and Tranco). These new .csv files are constructed according to Algorithm 1 and is done similarly for Tranco.

Algorithm 1 Get category domains

```
for date d in dates do
    for domain i in Cisco_d do
        if domain_i in SW then
            domainCSV_d ← (domain_i, rank_domain_i)
        end if
    end for
end for
```

Furthermore, we then wanted to extract each of the categories' ranks based on their domains. To accomplish this and get a better view of changes in trends we wrote another python script, simplified in Algorithm 2, which takes each domains' ranking from our new .csv files and adds to another final list with each category. Meaning each category gets a summation of scores from all domains within the given category, ranging from 1 to $\frac{1}{700}$. As an example, the fourth domain in our Cisco list will give the value $\frac{1}{4} = 0.25$ to the category corresponding to that domain in our final .csv dataset. As mentioned, this procedure is done once for each of our two different data sources.

After producing these .csv files, we end up with each category getting a relative value of it's popularity compared with each other and finally we used these category files to plot trends over our time period for our results and analysis, using a seven day moving average.

Algorithm 2 Get category rankings

```
for date d in dates do
    for domain i in domainCSV_d do
        for category c in categories do
            if domain_i in category_c then
                categoryCsv_d(c) += 1/i
            end if
        end for
    end for
end for
```

3 Results

In this section, the resulting relative ranks of each of the 14 categories, during the time period 2019-01-01 to 2022-03-31, is presented for both Cisco Umbrella and Tranco. The y-axis represents the calculated rank of the category, like described in section 2.4.

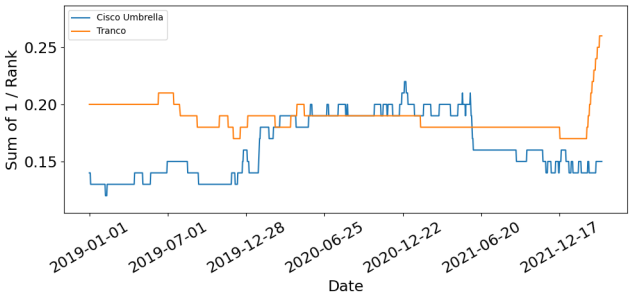

Figure 1. Adult

For the adult category, Cisco Umbrella shows a clear trend of an increase in popularity during COVID-19. Meanwhile, the data from Tranco show no such trend, but instead a sudden increase during the beginning of 2022.

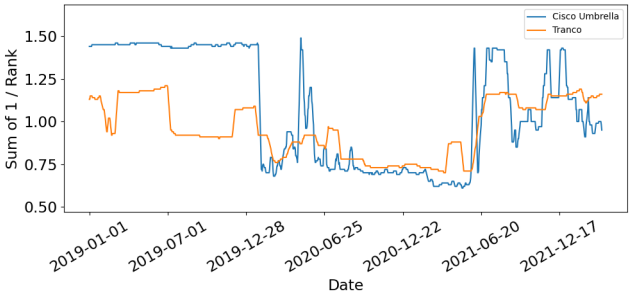

Figure 2. Arts and Entertainment

For the arts and entertainment category, both datasets suggests a decrease of popularity during COVID-19. However, the trend is more prominent for Cisco Umbrella.

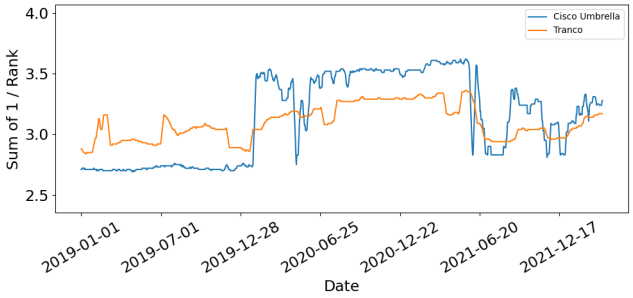

Figure 3. Computers, Electronics and Technology

For the computers, electronics and technology category, both datasets suggest an increase of popularity during COVID-19. In this graph we can see that the values are significantly higher than for the other categories. This is due to SimilarWeb's categorization where domains such as *google.com* and the most popular social media platsforms (*facebook.com*, *instagram.com* and *twitter.com*) all fall under this category.

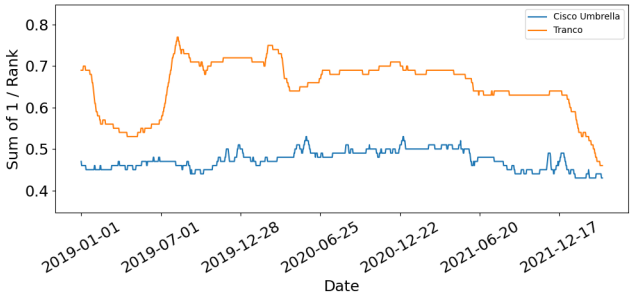

Figure 4. E-commerce and Shopping

For the E-commerce and shopping category, no clear trends related to COVID-19 is identified.

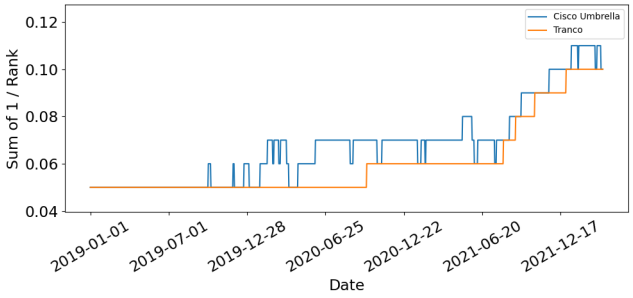

Figure 5. Gambling

For the gambling category, an increase in Cisco Umbrella can be seen during the beginning of COVID-19, only to drop of and settle on slightly higher popularity than pre-COVID-19, which is represented in both datasets. After COVID-19, gambling sees a significant increase in popularity.

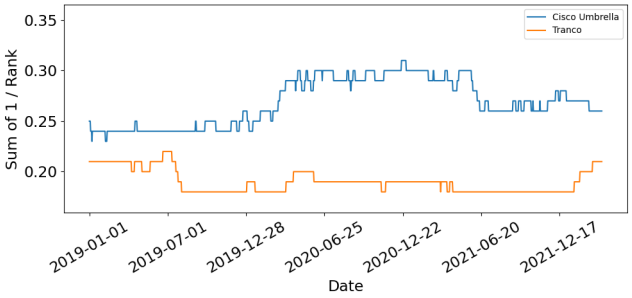

Figure 6. Games

For the gaming category, Cisco Umbrella suggests an increased and steady popularity during COVID-19, and post-COVID-19 it settles at higher level than pre-COVID-19. Tranco does not correlate with Cisco Umbrella, but instead show a slight increase during the onset of COVID-19, that soon tapers off.

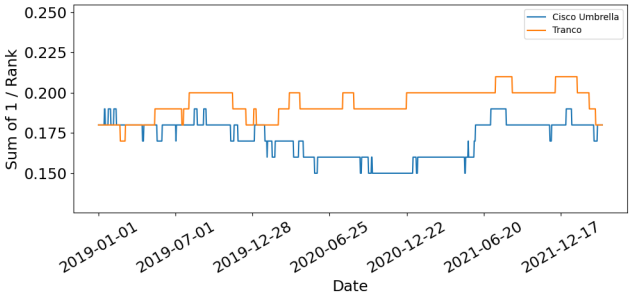

Figure 7. Health

For the health category, Cisco Umbrella show a slight decrease in popularity during COVID-19, which is reset post-COVID-19. Meanwhile, Tranco suggests a slight steady increase during most of the entire time period.

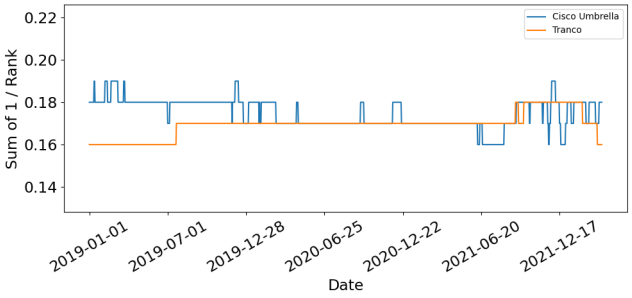

Figure 8. Hobbies and Leisure

For the hobbies and leisure category, Cisco Umbrella suggest a slight decrease in popularity during COVID-19, while Tranco does not show a clear trend related to COVID-19.

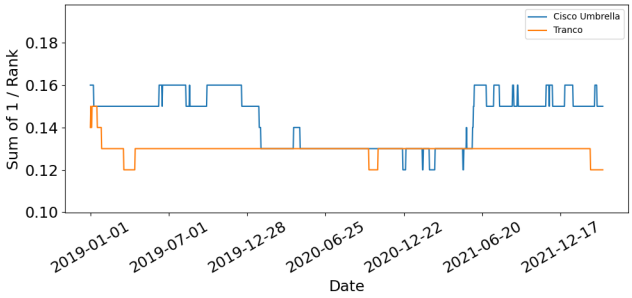

Figure 9. Home and Garden

For the home and garden category, Cisco Umbrella shows a clear trend of decreased popularity during COVID-19, that resets to pre-COVID-19 levels afterwards. Meanwhile, Tranco show no sign of popularity changes related to COVID-19.

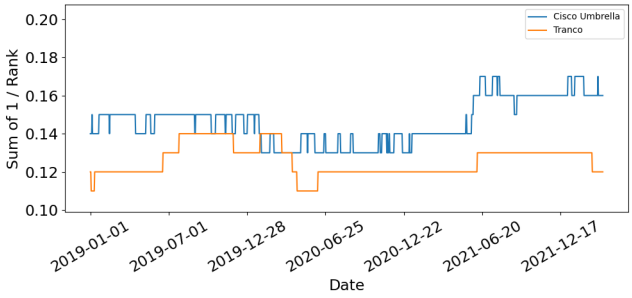

Figure 10. Jobs and Career

For the jobs and career category, both datasets suggests a decrease during COVID-19 and a slight increase after. The trends are more prominent for Cisco Umbrella.

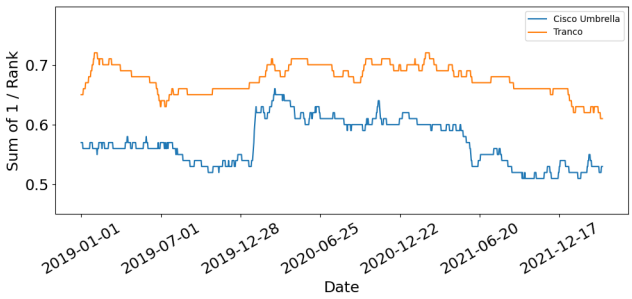

Figure 11. News and Media

For the news and media category, both datasets suggests an increase during COVID-19 and a reset to pre-COVID-19 levels after. The trends are again more prominent for Cisco Umbrella.

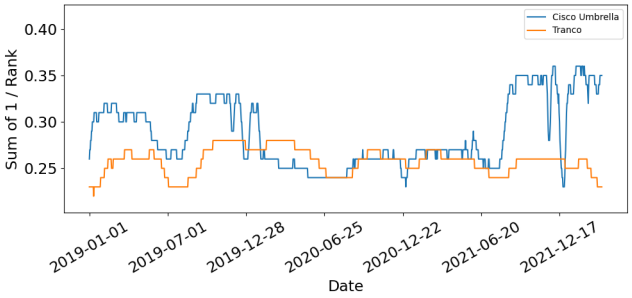

Figure 12. Science and Education

For the science and education category, Cisco Umbrella suggest a decrease in popularity during COVID-19, while no clear trend is shown for Tranco.

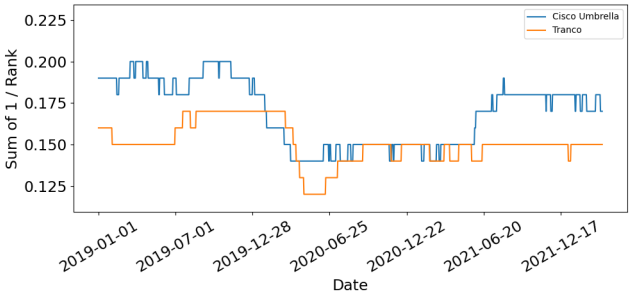

Figure 13. Sports

For both the sports category and travel and tourism category, both datasets suggests a significant decrease during COVID-19, especially during the beginning. However, the trend after COVID-19 differs, where Cisco Umbrella show an increase and Tranco show no change in popularity.

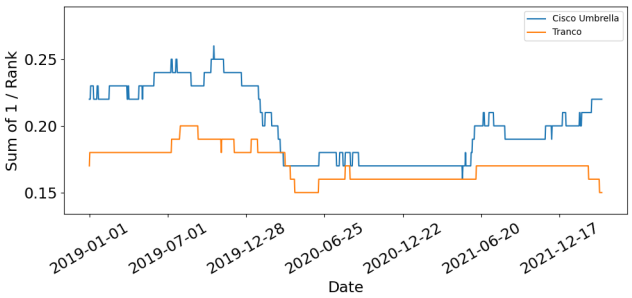

Figure 14. Travel and Tourism

4 Conclusion and Discussion

As seen in the figures above, results has been gathered in 14 different categories from both Cisco Umbrella and Tranco top list. One key aspect to note about these results is how respective lines representing Cisco Umbrella and Tranco differs from each other. Comparing the two, one can distinguish that Cisco Umbrella tends be more sensitive to daily change while Tranco is more stable. This sensitivity can be an effect of how Cisco Umbrella tracks subdomains while Tranco does not since it is more plausible for several subdomains, all with the same top domain, to all lower their ranking by a small amount on a day than it is for a top domain to lower it's ranking by a large amount.

Another observation to note is the data sets varies from day to day even though their trend seem similar. This gives confidence that trends from the graphs are more likely to reflect the reality than the rankings from one of the sets. This highlights the importance to have several different data sets to compare since they do not concur with each other in every aspect, and can therefore be used to skew the agenda in one direction.

As we can see, most of the graphs lies within one of two groups in their trend. They either have an temporary increase or decrease during the pandemic, and later go back to somewhat of a similar level like before. In the decrease group we can see categories that mostly rely on human interaction, or were limited by the pandemic in some way, like "Arts and Entertainment", "Travel and Tourism" or "Home and Garden". On the contrary, we see categories that are often done in solitude for the increasing group. This includes "Adult", "Games" and "News and Media" among others. There are two outliers among the categories, that don't follow either of the two trends, that is "E-commerce and Shopping", and "Gambling". "E-commerce and Shopping" instead has the same numbers across the whole period, with the slight notion of an increase during the pandemic. "Gambling" has the same trend, but more dominant. Here we can also see that the trend does not go back to normal for the period after the pandemic, but keeps increasing in a steady pace.

Finally, we may conclude the report stating that our DNS traffic analysis clearly suggests changes in web trends during COVID-19, and that several trends correlated among our two sources.

References

[1] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. *Proceedings 2019 Network and Distributed System Security Symposium* (2019). https://doi.org/10.14722/ndss.2019.23386

[2] Quirin Scheitle. 2022. *Top Lists.* https://toplists.github.io/

[3] Similarweb. 2022. *Top Websites Ranking.* https://www.similarweb.com/top-websites/

[4] Tranco. 2022. *Tranco.* https://tranco-list.eu

[5] Cisco Umbrella. 2022. *Cisco Umbrella.* https://umbrella.cisco.com/