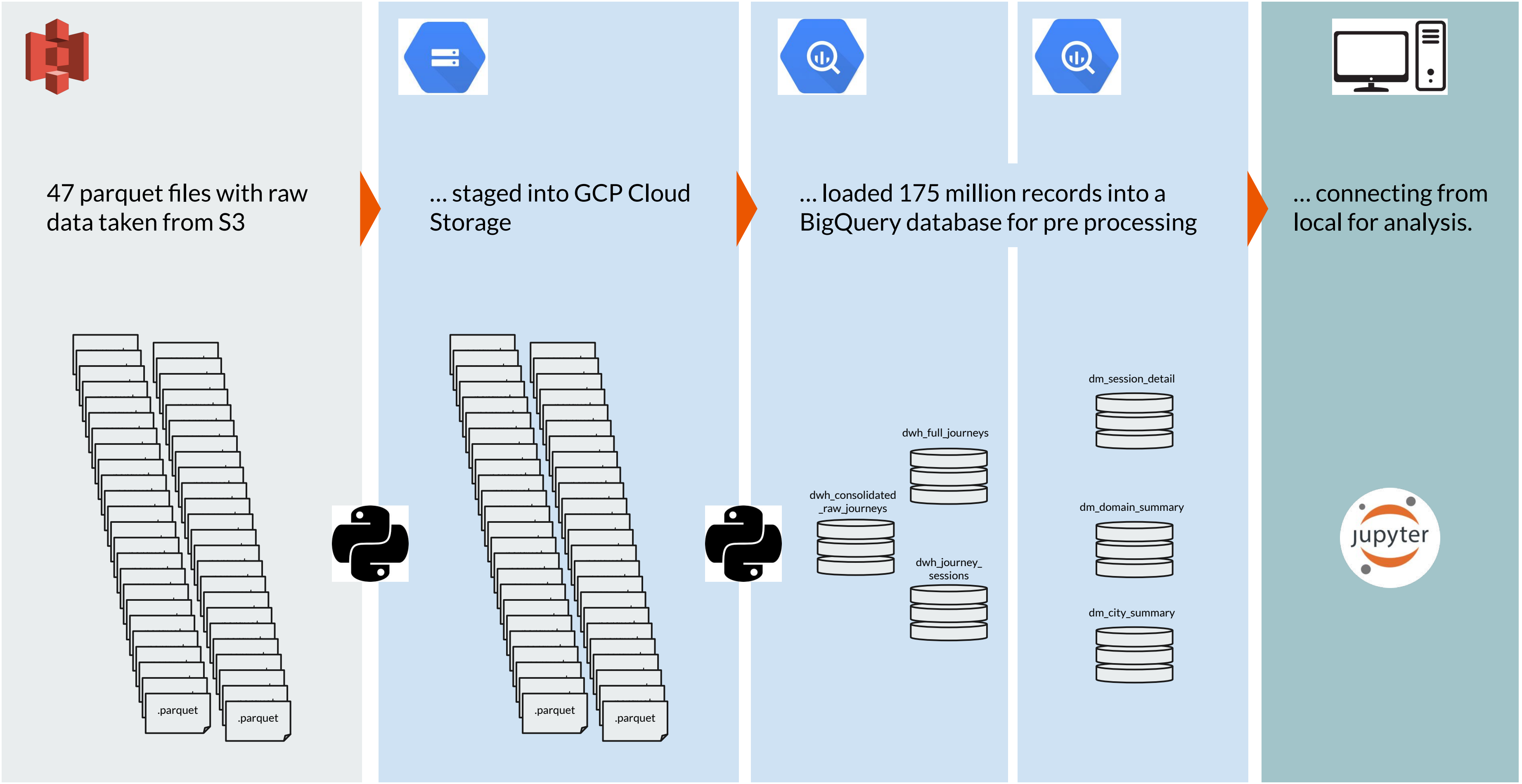
Journey to TripAdvisor

Adán Tinoco Marquina

Oct’23

- Given the size of the files, and the available computing power, a Python script was used to transfer the data into Google Cloud to build an architecture that could withstand the requirements.
- To use all the data, a database was designed with three layers:
    - Staging zone for raw files.
    - Warehouse integrating files into a single table, and basic transformation tables.
    - Data mart with processed tables for analysis purposes.
- The analysis was then carried out in local environment, using a Jupyter Notebook.
- Codes and further detail can be found in:
    - https://github.com/adanttmm/JourneyToTripAdvisor.



47 parquet files with raw data taken from S3

... staged into GCP Cloud Storage

... loaded 175 million records into a BigQuery database for pre processing

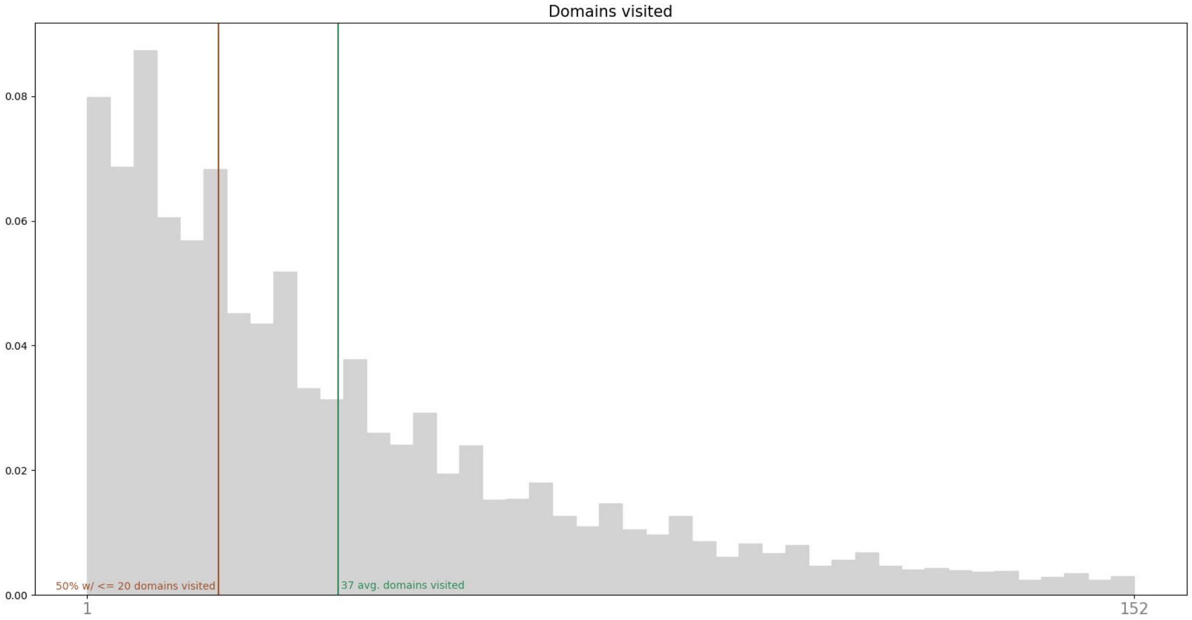
... connecting from local for analysis.

- After loading over **175M rows** into BigQuery, the suggested table with a row per pageview is created.
- Column 'useragent' wasn't found in raw files.
- Strange values arise when observing timestamp lag differences.
- Using BigQuery's TIMESTAMP_SECONDS function (seconds count after 1970-01-01 00:00:00 UTC) shows mismatch between date and transformed timestamp.
- Padding timestamps to 10 digits **18M** show difference on date and timestamp, and only 41K have differences greater than 1 day.
- Padded timestamp are used to get the event time in DHW and the 'eventdate' timestamp is used for the **41K atypical values**.
- Rows with NULL 'userid', 'eventdate', or 'eventtimestamp' are removed.
- Column 'countrycode' appears to be mislabeled from origin, and in fact carries 'useragent' information, so column name will be changed in DWH tables.
- After cleaning the data, creating the ranked url succession, and fixing the timestamp issues, **123M records remain** which account for 70% of the original volume.
- There's an **15 mins. average per page with 69 average pages per user**, which looking at the percentiles (5% buckets) prove highly skewed by atypical values.
- Moreover **only 5% exceed 10 mins between pages;** given this and the industry standard of 30 mins. between events, journeys are broken into sessions in ta new DWH dataset.

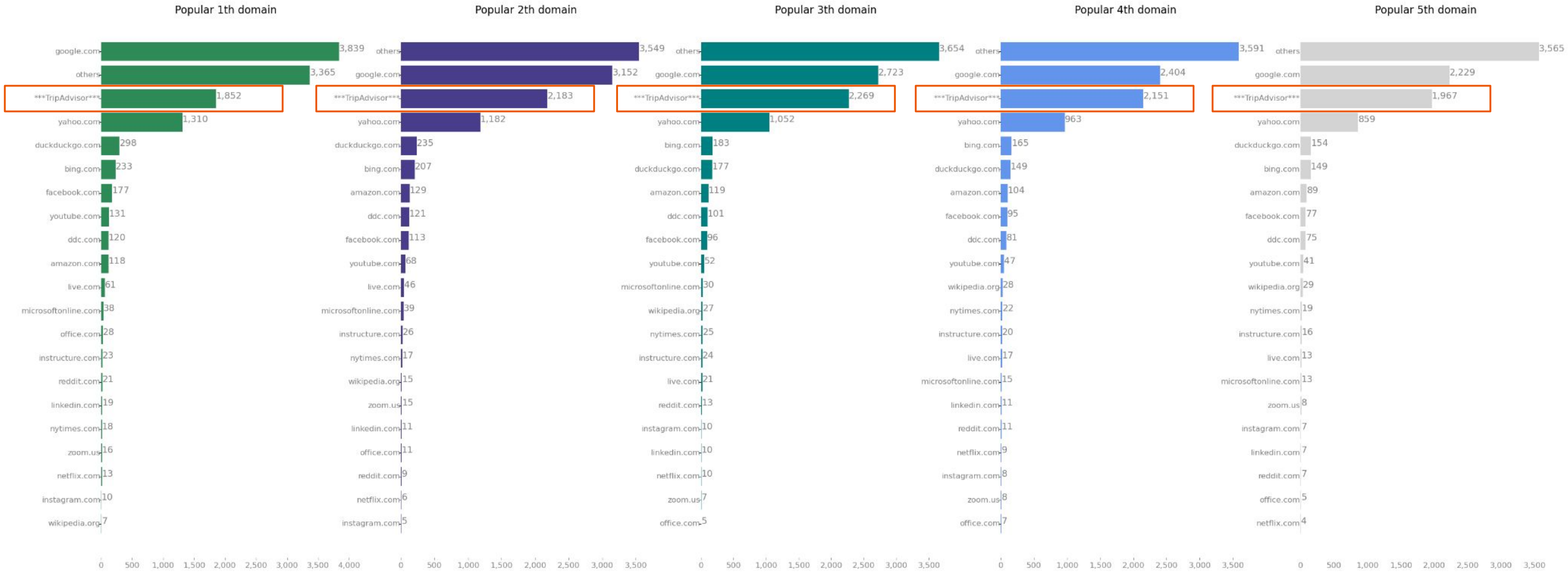
Task 1: Understanding the journey

- With the session definition previously explained, each user shows **519 average sessions over 3 year.**
- The average time spent per url is a **minute and a half**, but the distribution shows a very long tail with atypical values.
- There are **1.5M urls** visited in these sessions, corresponding to **958K domains**. Domains are used for analysis purposes.
- Out of the **9M** observed sessions, **26K** go to or through TripAdvisor.

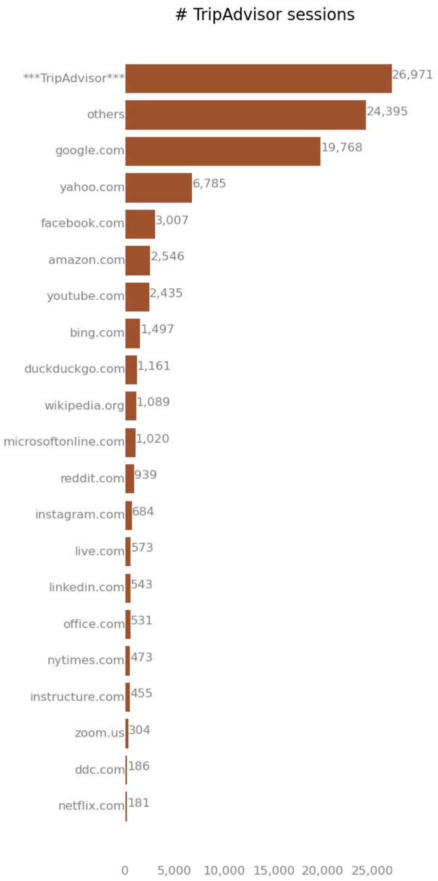
TripAdvisor sessions are almost 4 times larger than the rest with 37 domains visited after removing atypical values.



TripAdvisor domains appear in the top 3 during the first steps of the journey, signaling that the users look it up from the beginning of their journey and exploring further after it.

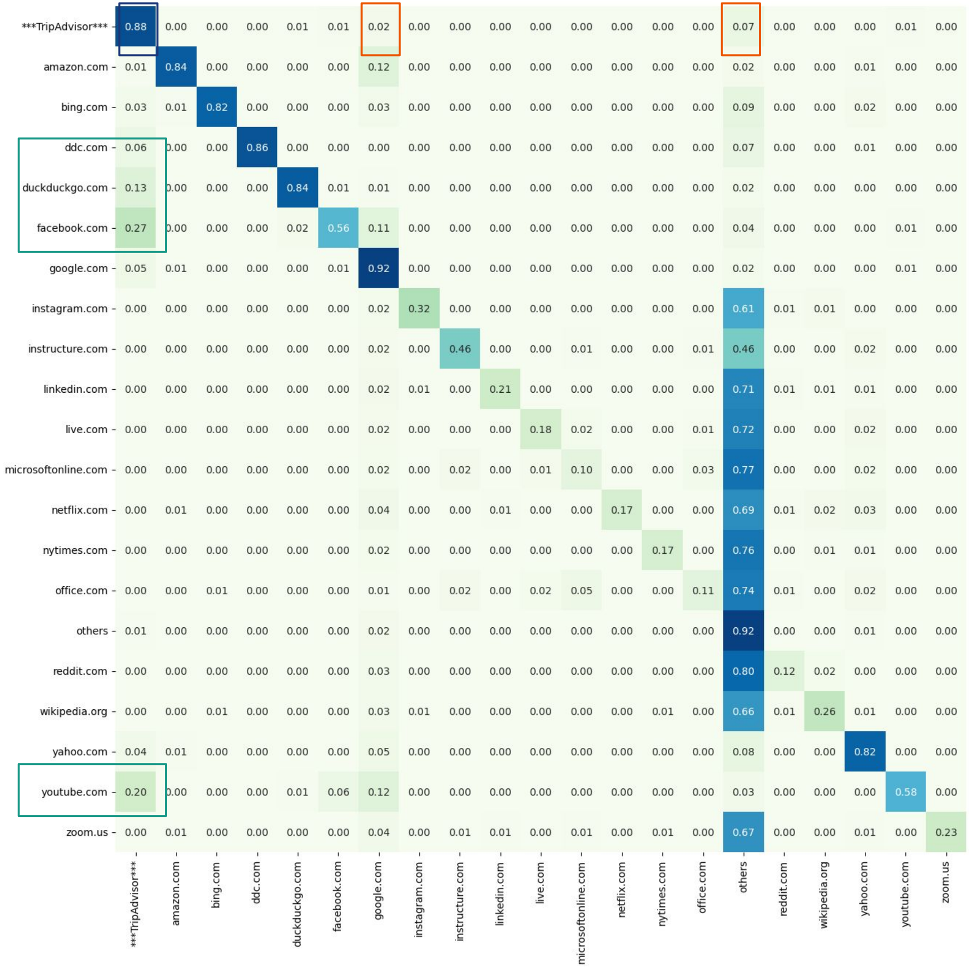


- Additional to TripAdvisor’s, these sessions visit:
    a. **Google, Yahoo, Bing, and DuckDuckGo signaling a research behaviour** due to the search engines of these domains, reinforced by popular **Wikipedia** visits.
    b. **Amazon** also appear in the top domains, showing a **purchasing propensity** for these sessions.
    c. **Social networks** appear as well, possibly product of **digital campaigns**.



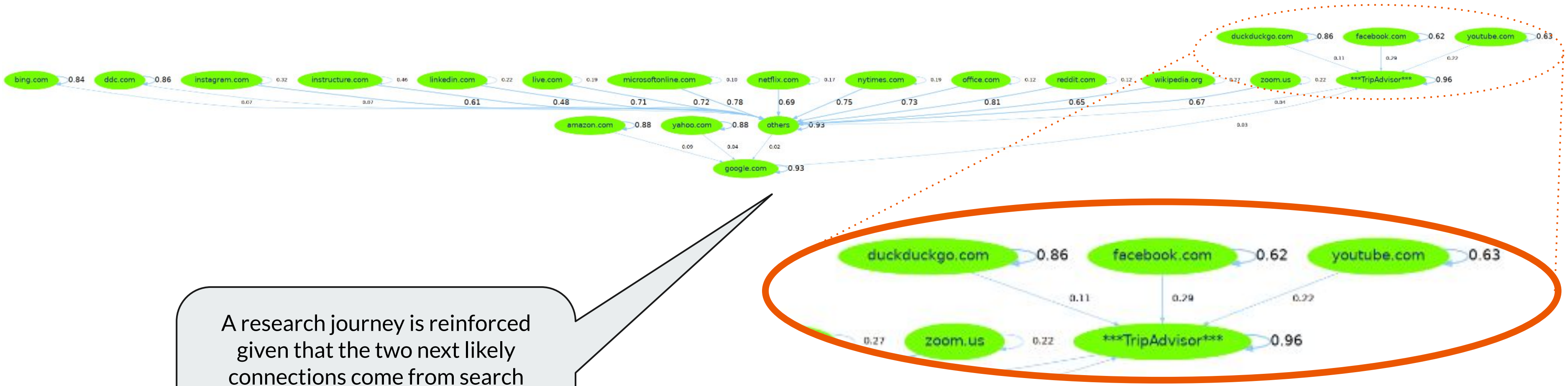
Looking at the transition probability for the journeys that go through TripAdvisor general patterns arise:

- Most likely **sources leading to TripAdvisor** are Facebook, Youtube, Google, DuckDuckGo, and DDC.com.
- Once in our domains there's an **80% probability that the journey will continue within.**
- The most likely next steps **going out of TripAdvisor is Google**.



	TripAdvisor	amazon.com	bing.com	ddc.com	duckduckgo.com	facebook.com	google.com	instagram.com	instructure.com	linkedin.com	live.com	microsoftonline.com	netflix.com	nytimes.com	office.com	others	reddit.com	wikipedia.org	yahoo.com	youtube.com	zoom.us
TripAdvisor	0.88	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.01	0.00
amazon.com	0.01	0.84	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00
bing.com	0.03	0.01	0.82	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.02	0.00	0.00
ddc.com	0.06	0.00	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.01	0.00	0.00
duckduckgo.com	0.13	0.00	0.00	0.00	0.84	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
facebook.com	0.27	0.00	0.00	0.00	0.02	0.56	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.01	0.00
google.com	0.05	0.01	0.00	0.00	0.00	0.01	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.00
instagram.com	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.01	0.01	0.00	0.00	0.00
instructure.com	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.46	0.00	0.00	0.01	0.00	0.00	0.01	0.46	0.00	0.00	0.02	0.00	0.00
linkedin.com	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.21	0.00	0.00	0.00	0.00	0.00	0.71	0.01	0.01	0.01	0.00	0.00
live.com	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.18	0.02	0.00	0.00	0.01	0.72	0.00	0.00	0.02	0.00	0.00
microsoftonline.com	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.01	0.10	0.00	0.00	0.03	0.77	0.00	0.00	0.02	0.00	0.00
netflix.com	0.00	0.01	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.01	0.00	0.00	0.17	0.00	0.00	0.69	0.01	0.02	0.03	0.00	0.00
nytimes.com	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.76	0.00	0.01	0.01	0.00	0.00
office.com	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.02	0.05	0.00	0.00	0.11	0.74	0.01	0.00	0.02	0.00	0.00
others	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.01	0.00	0.00
reddit.com	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.12	0.02	0.00	0.00	0.00
wikipedia.org	0.00	0.00	0.01	0.00	0.00	0.00	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.66	0.01	0.26	0.01	0.00	0.00
yahoo.com	0.04	0.01	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.82	0.00	0.00
youtube.com	0.20	0.00	0.00	0.00	0.01	0.06	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.58	0.00
zoom.us	0.00	0.01	0.00	0.00	0.00	0.00	0.04	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.67	0.00	0.00	0.01	0.00	0.23

The most probable connection going into TripAdvisor comes from Facebook domains, which may be an indicator of the success of digital campaigns.

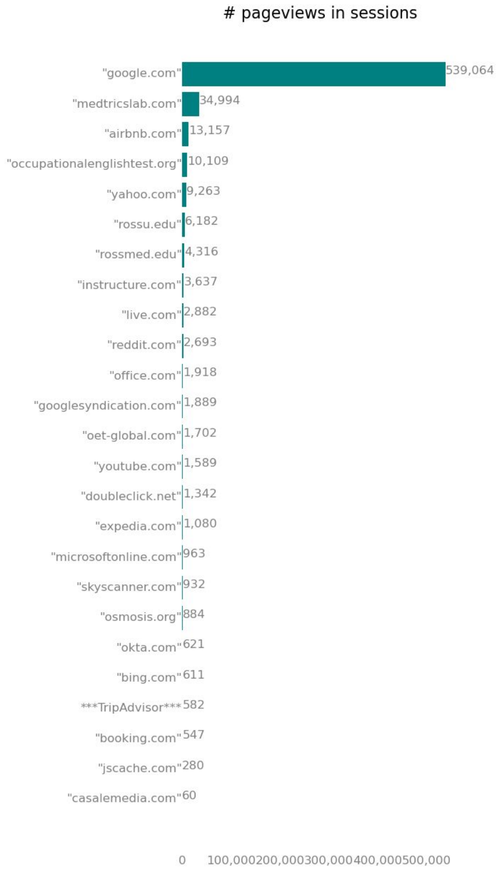


A research journey is reinforced given that the two next likely connections come from search engines, and Google accumulates the biggest probability going out as a single domain.

Task 2: Longest TripAdvisor journey

The longest session going through TripAdvisor lasted 111.9 days of continuous browsing, visiting 641,297 urls.

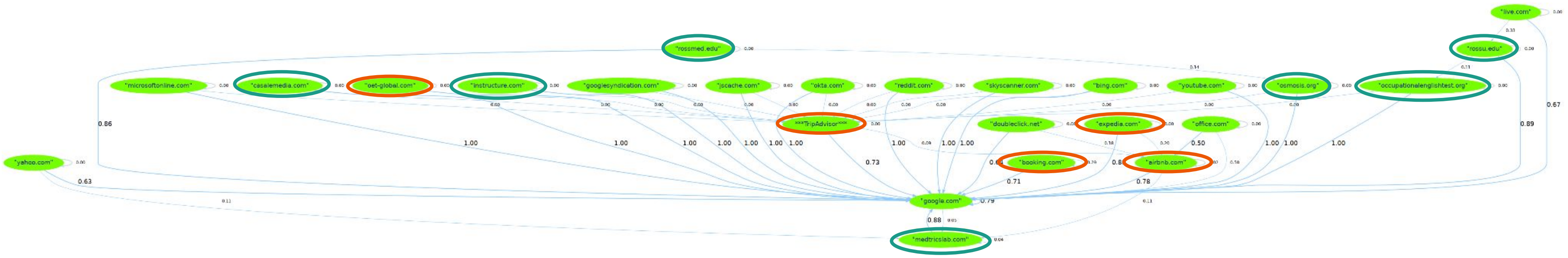
These pattern is very **unlikely to come from a single individual.**



* Most domains visited are either **travel**, search engines and **educational websites**.

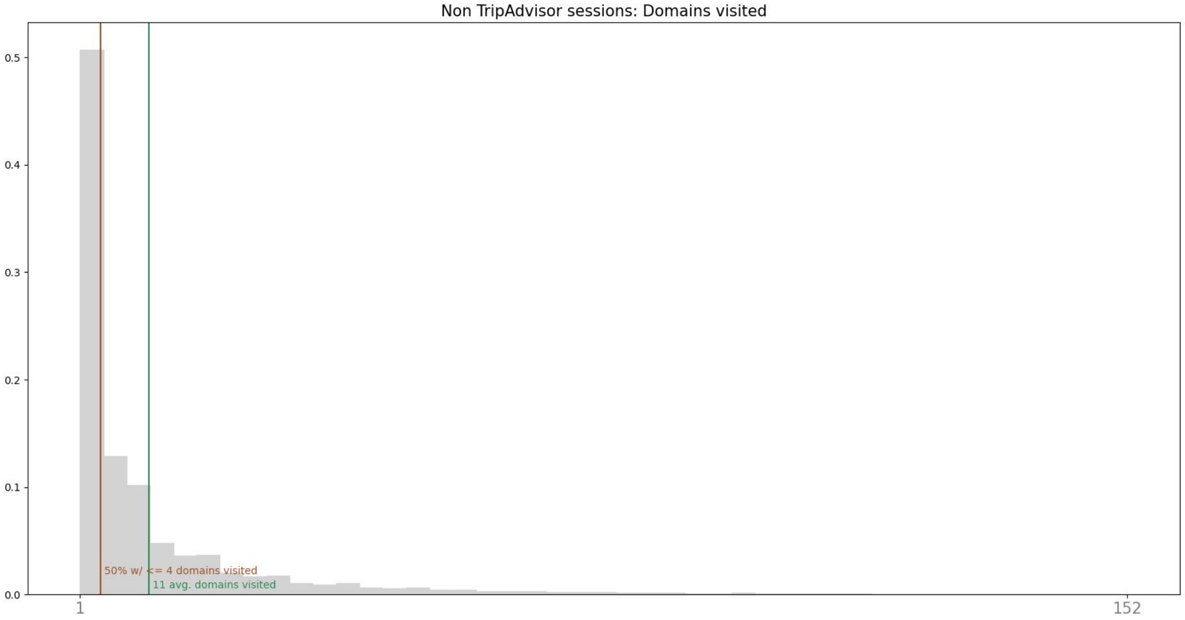
* Given that this session comes from a single computer, traffic is likely to come from multiple individuals using a shared resource.

* Looking into the type of educational platforms and travel research pattern shown, it's likely to be a resource from a medical education institution, with people traveling due to their activities.

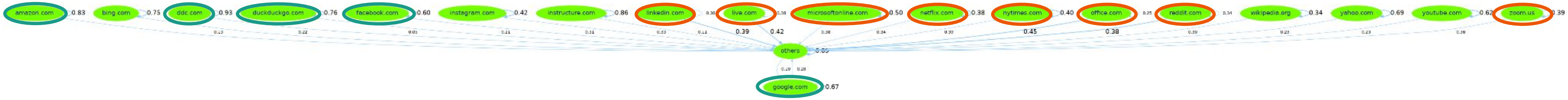


Task 3: Engaging journeys

Sessions that don't go through TripAdvisor show a shorter journey, signaling that these users have a fixed purpose.



- Looking into the journeys that don't go through TripAdvisor, most domains are search engines and social network.
- **Less likely sources of traffic** for TripAdvisor should be avoided, given that the behaviour displayed shows a straightforward path to a specific need, like work or entertainment.
- Domains that are **proven sources** can be leveraged with further efforts to engage people.


amazon.com
0.83
bing.com
0.75
ddc.com
0.93
duckduckgo.com
0.76
facebook.com
0.60
instagram.com
0.42
instructure.com
0.86
linkedin.com
0.30
live.com
0.30
microsoftonline.com
0.50
netflix.com
0.38
nytimes.com
0.40
office.com
0.25
reddit.com
0.34
wikipedia.org
0.34
yahoo.com
0.69
youtube.com
0.62
zoom.us
0.39
0.13
0.22
0.05
0.21
0.31
0.33
0.11
0.39
0.42
0.38
0.34
0.33
0.45
0.38
0.30
0.23
0.23
0.36
others
0.05
0.10
0.28
google.com
0.67