Covid-19 Attention

A study on the indirect impact of a crisis through information seeking behaviour on Wikipedia

Roman Dahm
VU
Amsterdam, the Netherlands
contact@romnn.com

Yannick Brunink
VU
Amsterdam, the Netherlands
yannickbrunink@gmail.com

1. INTRODUCTION

The Coronavirus pandemic (COVID-19) is not the first pandemic with a worldwide impact. Yet, it is the first one in a time of widespread internet access. This enables researchers to quantify not only the direct impact [1] but also the indirect impact on human lives, such as human needs, interests and concerns. Recent studies suggest that many of the challenges of the pandemic are concerning this indirect impact [6] and should be incorporated in the decision making process[10] .

One illustration of the above mentioned widespread internet access is the online encyclopedia WIKIPEDIA [2], which is free and easily accessible for anybody with an internet connection. WIKIPEDIA keeps track of every page visit and reports daily aggregates publicly.

The starting point of this study is to capture the shifts in attention by calculating the cumulative page-view differences for different topics around selected *change-points* with respect to COVID-19 measures.

Although the name COVID-19 might suggest otherwise, the COVID-19 pandemic started to have an impact on peoples lives starting from 2020. Therefore, we will consider 2019 as a baseline year and focus on possible shifts in attention in 2020 and 2021. The raw volume attention shift analysis will present a first glance of the impact on COVID-19 information seeking patterns in general. Subsequently, we will target specific events and dates, referred to as *change-points*, to investigate in more detail. By grouping page-views per topic, we are then able to apply difference-in-difference regression for time windows pre- and post-change-point. Thereby we strive to capture what goes on in people's minds during these times.

The change-points are specific to a country and include few manually pre-selected dates as well as change-points estimated from the *stringency-index* (SI) of a country. The SI is a scale to measure the severeness of measurements, provided by the University of Oxford [3].

By applying the above-mentioned difference-in-difference regression for change-points per country we can investigate the indirect impact of COVID-19 through the SI proxy by investigating the nature of topics with the most pronounced shift in attention.

To summarize, the goal of this study is to quantify the indirect impact of the pandemic by entering the mind of internet users world-wide via their information-seeking patterns on WIKIPEDIA and thereby capturing shifts in attention. By linking the latter, either in volume or per topic, to the severity of the COVID-19 measures as well as presenting an overview of what topics and pages are of interest during COVID-19 related events we thrive to complete the picture of the indirect impact of the COVID-19 pandemic.

In the next section, we give an overview of related work of attention shifts during pandemics and an introduction of the algorithms used in this study. Subsequently, we will present our Research Questions and project setup. Finally, we discuss our results and conclusions.

All code that was used in this study has been made available as open-source[4]. Furthermore, an interactive data visualization with results obtained in this study is available at `https://romnn.github.io/lsde2021/`.

2. THEORETICAL FRAMEWORK

2.1 Related work

In this section we review related work that studied the effect of user behaviour during a virus outbreak. Firstly, Ribeiro et. al (2020)[9] present a longitudinal analysis of WIKIPEDIA page-view data during the COVID-19 pandemic. Their analysis was three-fold (1) shifts in overall page-count volume. (2) Shifts in Information seeking patterns. (3) Shifts in Topic-specific page-count volumes. They conclude that the sharp decrease in human mobility has boosted the

[1] COVID-19: 219M Cases, 4.55M Deaths as of October 2021
[2] `https://www.wikipedia.org/`

[3] `https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker`
[4] `https://github.com/romnn/lsde2021`

Proceedings of the VLDB Endowment, Vol. , No.
ISSN 2150-8097.
DOI:

volume of information seeking in general and has changed the nature of information sought. Furthermore, once mobility returned to pre-COVID-19 levels, the volume did as well, yet the nature of the content did not. Finally, some topics such as *Biology* saw brief increases, whereas others had persistent increase (*Video Games*) or decrease (*Sports*).

Panisson et. al (2020)[8] investigated the impact of news exposure on collective attention in the United States during the 2016 Zika epidemic. Although their main research question is not directly related to this study, one of their findings is. Namely, they state that – contrary to their hypothesis, WIKIPEDIA visiting patterns (in terms of Zika related pageviews) were not significantly correlated with the magnitude or extent of the epidemic. Although one has to keep in the mind that the Zika virus was labeled an epidemic, and thus was of a different scale compared to the world-wide impact and media coverage of the COVID pandemic.

Abay and Tafere (2020)[5] studied the indirect impact of the COVID-19 pandemic with Google-search data. They found that the decrease in demand for certain services is mostly driven by social distancing and lockdown measures. Furthermore, they observed a relationship between the magnitude in changes of demand and the government responses to the pandemic.

Ribeiro et. al (2020) performed a similar study, however their study comprised of a shift in attention from 2019 to 2020, whereas we estimate the effect for the years 2020 and 2021 compared to the base year 2019. Panisson et. al(2020) and Abay and Tafare(2020) observed relevant relationships, however those studies were of different nature in scale and data. It is therefore we argue that this study will contribute in completing the picture of the indirect impact of the COVID-19 health crisis.

2.2 Statistical Models & Algorithms

2.2.1 *Difference-in-Differences*

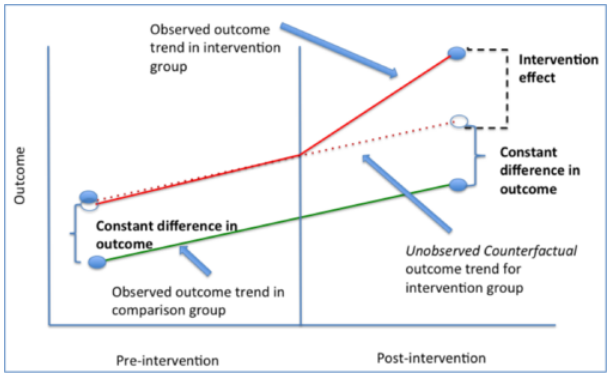
Difference-in-differences is a statistical technique used to estimate treatment effects by comparing the change in the differences in observed outcomes between treatment and control groups, across pre-treatment and post-treatment periods. The difference-in-differences method aims at separating the true treatment effect from simultaneous changes that would have occurred even without the treatment [9], as illustrated in figure 1. Naturally, one can not withhold the impact of COVID-19 on half of the population. Therefore, we use 2019 as the control group and 2020 to 2021 as the treatment group.

Subsequently, the treatment effect can be estimated by subtracting the average change in the control year from the average change in the respective COVID year. For both years, two time periods are required. Before and after the event (treatment). More formally, the regression model can be defined by formula 1.

$$y_{i_t} = \beta_0 + \beta_1 P_t + \beta_2 T_i + \beta_3 (P_t \cdot T_i) + u_{i_t} \quad (1)$$

$$\hat{\beta}_3 = (\overline{y_{T=1,P=0}} - \overline{y_{T=0,P=1}}) - (\overline{y_{T=0,P=1}} - \overline{y_{T=0,P=0}}) \quad (2)$$

Where y is the outcome of interest, P is a dummy variable for the second time period, and T is a dummy variable for the treatment group. The interaction term $P \cdot T$ is equivalent to a dummy variable of 1 for observations in the treatment group in the second period.



Source: www.publichealth.columbia.edu

Figure 1: Difference-in-Difference on events.

In order for the difference-in-difference estimator to hold, the parallel trend assumption is a requirement. Since 2021 is relatively far from 2018 one can argue that this assumption is no longer valid. Another possibility is to take the previous year as base year [9], however, we argue that comparing 2019 with 2020 is not relevant as both years have been similarly affected by the COVID-19 pandemic. We hope to obtain a more precise estimate by comparing against the pre-pandemic baseline year 2019. However, the possible violation of this assumption should be kept in mind when interpreting the results and comparing with the absolute difference approach presented in the next section.

2.2.2 *Absolute-Difference*

In addition to the difference-in-difference regression approach described in the previous chapter, we ranked shifts in attention per topic based on their absolute difference of the means in the time window pre- and post-change-point. This approach is simpler and more susceptible to seasonal trends that can have a rapid influence over a short window size (e.g. the topic *Christmas* for the change-point on the day of Christmas. On the other hand, it naturally handles new emerging topics better than the difference-in-difference approach, which assumes that topics are always present and only vary in their popularity.

3. RESEARCH QUESTIONS

In order to be able to present a more complete image on the indirect impact of COVID-19 on attention shift we have formulated the following research questions.

- **RQ1** Does COVID-19 and its measures have an impact on information seeking patterns? (what information is looked up how much?)
- **RQ2** Are those trends subject to temporal or spatial differences?

For RQ1 we do expect to see an impact on information seeking patterns. In the beginning of the pandemic (January 2020), we expect to see an attention shift towards virus related pages. Furthermore, we hypothesise that when the stringency level is high the nature of information seeking

patterns is different. More precisely, there will be a shift towards topics that relate to entertainment, gaming, indoor activities, or self education about broad topics. On the other hand, topics related to outdoor activities, public places, or cultural activities such as music concerts will decrease. Although those expected results might be intuitive, they do provide possible evidence of the fact that human-behaviour does actually change during high-stringency periods. Thus, for RQ-2, we hypothesise that there will be temporal and spatial differences in these patterns, namely if one country's stringency index is low, whereas in another country it is high we do expect to see this in the information seeking patterns.

3.1 Technological Research

As we expected, we did not face considerable technological problems, as a lot of good tooling for HTTP downloading, change-point detection (4.3), distributed data processing, and interactive web visualization is readily available. However, working with a dataset on the terabyte-scale was unfamiliar, as many popular tools are not able to work with data on that scale. Hence, using Apache Spark[11] was necessary to be able to process the data efficiently in parallel. The latter allowed us to both download data in parallel without rate-limiting as well as process large amounts of data that did not fit into main memory in parallel. Hence, we did not identify any scalability issues in that regard. However, as to be discussed in Section 4.4, working with large graphs is not possible using traditional data frames. In our case it was possible to fit the category graph into main memory, however, for even larger problem sizes, one must consider one of the many available options for distributed graph processing such as GraphX for Spark[13].

4. PROJECT SETUP

4.1 Dataset Statistics

Central to our data analysis are the daily pageview statistics[5] provided by the Wikimedia Foundation [6], which contain the total number of `views per day` per `page ID` on a `wiki` (e.g. *de.wikipedia*) by their `device type` (e.g. *mobile* or *desktop*). Wikimedia provides three versions of the page-view data, *user*, *spider*, and *automated*, distinguishing between the origin of the traffic. Naturally, we use the user dataset, for which an example entry is given in Table 1.

Wiki	de.wikipedia
Page Title	1._Weihnachtsfeiertag
Page ID	9075
Device	mobile-web
Total views	1
Hourly views	l1

Table 1: Sample entry from the daily page-view dataset. (`pageviews-20180101-user.csv`, line 3549560)

To get an overview of the data, we performed an initial exploratory analysis on 12 random samples of daily page-view data. On average, each daily page-view dataset contains roughly 22.7 million rows. Equivalently, 22.7 million pages are viewed per day, with a standard deviation of about 691 thousand.

[5] `https://dumps.wikimedia.org/other/pageview_complete/`
[6] `https://www.wikimedia.org/`

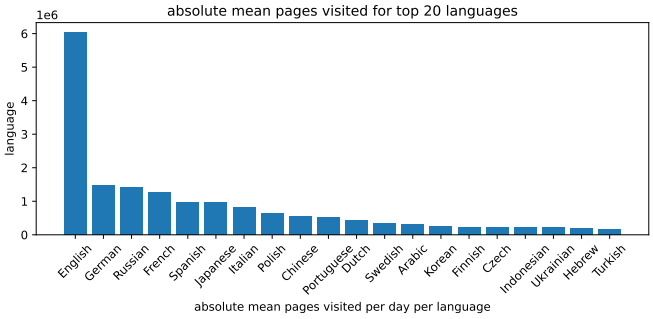


Figure 2: Mean absolute number of pages visited daily for the 20 most popular languages.

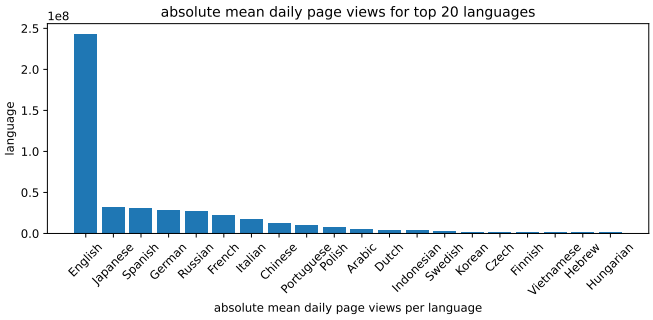


Figure 3: Mean absolute number of daily page views for the 20 most popular languages.

First of all, it can be observed that the data distribution is highly skewed, as is to be expected. The top 100 pages (0.00016%) account for 4.52% of all page views, while 98.75% of pages have less than 100 views per day. The top 100.000 pages (0.1622%) account for 32.97% of all page views. On average, each page is viewed 9.1 times per day, with a high standard deviation of 1984.75. Of the selected samples, 13% of pages can be found in each sample and 22% can be found in at least 75% of them, hinting on the fact that there naturally is considerable overlap between daily page-view data, but also a large variation in the pages viewed.

Furthermore, as hypothesized, the English Wikipedia accounts for a vast majority of all pages and page views, as can be seen in Figures 2 and 3.

To complement the page-view data, we also consider the official Wikimedia SQL database dumps[7], which provide useful tables such as the `langlinks` table, to map pages to different languages, as well as the `category` and `categorylinks` table to map pages to categories. The latter SQL dumps are provided separately for each language, however, it is sufficient to download just the latest versions and just the English `category` and `categorylinks` tables.

Furthermore, we consider the stringency index provided by the University of Oxford [7], a measure for the severeness of COVID policies in a country for 186 countries over a duration from 2020 to October 2021. For illustration purposes, the stringency index for Germany is shown in Figure 4.

4.2 Pre-Processing and Data Pipeline

[7] `https://dumps.wikimedia.org/dewiki/20211001/`

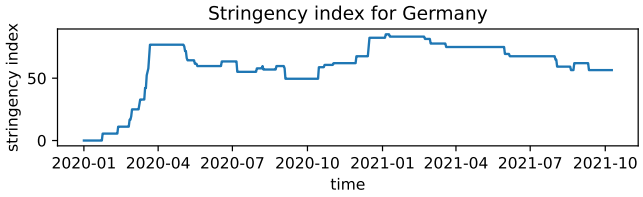


Figure 4: Stringency Index for Germany from 2020 to 2021.

We used multiple SPARK worker nodes to download daily page-view CSV files as well as the SQL dumps in parallel. Subsequently, self contained and optimized Parquet files of the daily page-view data were created as follows:

1. Page-views were grouped on their *wiki* and *page ID* and all views from different *devices* were summed.

2. Using the inter-language links provided in the `langlinks` database table, we included the English *page ID* for each page and dropped pages for which no English translation existed. This serves as a way for us to validate and interpret any findings for languages we do not speak and allows us to identify trends in topic popularity across different languages.

3. Among the total 323 languages[8] of WIKIPEDIA, we chose 41 languages[9] which we were able to clearly map to countries with sufficient population size for which data on stringency and COVID-19 infections is available. Initially, we planned to use fewer languages and exclude languages such as Chinese which we cannot understand and validate, however, when considering only articles for which we know the English translation, we decided to also investigate changes in attention for other languages. The page-view dataset is joined with language related metadata such as the language code, language name, country, *ISO3* country code, as well as population size.

4. Based on the English *page ID* the page-view dataset is joined with the English article dataset containing mapped topics of different granularity's. Section 4.4 will provide more detail on how topics were mapped.

Using the optimized Parquet page-view dataset, a yearly dataset was aggregated and used to create yet another dataset with page-views grouped by topics.

For all operations on CSV or Parquet datasets, SPARK *data frames* were used to distribute the processing on to multiple cores and machines. Wherever possible, intermediate results were stored to build resumeable pipelines. In this process, we frequently used partitioning on the wiki language to be able to process or load different languages in parallel.

[8] `https://commons.wikimedia.org/w/api.php?action=sitematrix&smtype=language&format=json`

[9] A total of 51 languages were chosen, some of which – such as Chinese, Gan Chinese, Min Dong Chinese, Classical Chinese, and Chinese (Min Nan) are aggregated to produce 41 final languages.

4.3 Change Point Detection

To detect *change-points* in the stringency index of a country, a dynamic programming approach was used[10].

Given a target number of change-points k, it finds the exact minimum of the sum of costs by computing the cost of all sub-sequences of a time series using an algorithmic dynamic programming algorithm. The complexity of finding k change-points in a series of n samples is $\mathcal{O}(kn^2)$ and finds an optimal solution. We found that for the discrete stringency signal, $k = 10$ for a period of almost two years ($n < 730$) resulted in all significant change points being detected without introducing too much noise.

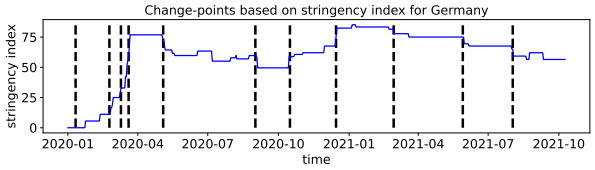


Figure 5: Detected change-points for stringency index of Germany.

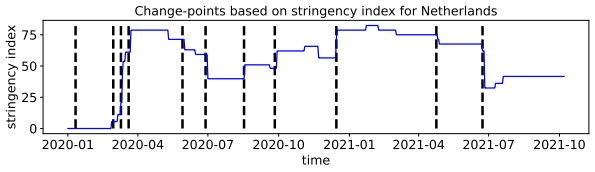


Figure 6: Detected change-points for stringency index of the Netherlands.

After the automatic change point detection, we manually added an important global event, namely the the first confirmed death in China on January 11th 2020[12]. Samples of the detected change-points for Germany and the Netherlands are shown in Figures 5 and 6 respectively.

4.4 Categorization

To be able to identify broad changes in topic attention across the large set of WIKIPEDIA articles, categorization of articles to high level topics is necessary. Most of the roughly 6 million English articles are categorized to very fine grained categories, most of which including the name of the article itself.

The Mediawiki Foundation provides the natural language processing model (NLP) API service ORES[3] for querying high level, coarse grained *topics* of an article revision. However, due to heavy rate-limiting and a very limited set of only 51 total categories (e.g. *Media*, *Sports*) with primary categories *Culture*, *Geography*, *History and Society*, and *STEM*[2], we used a heuristic algorithmic approach that recursively uses the WIKIPEDIA categories and category pages of category pages to build a *category-graph* that can be used to search for parent categories of different granularity.

Given the category graph $G = (V, E)$ of nodes V and directed edges E, where leaf nodes represent articles, inner nodes represent categories, and edges represent an is-in-category

[10] The *Dynp* implementation of the *ruptures*[4] python library for change-point detection was used.

relation between two category nodes or an article and a category node, we find higher level categorization for an article using a directed, depth-limited, level-based breadth-first-search (BFS) approach. The graph is constructed using the supplementary English WIKIPEDIA `categorylinks` and `articles` SQL tables, which were parsed using a hand written parser. The resulting category graph consists of 8 million nodes and 75 million edges, which is further reduced to 39 million (51%) edges after removing about 30 thousand hidden categories[11] nodes. The average node out degree is 5.4 for leaf article nodes and 2.95 for inner category nodes.

Starting a BFS from a given article leaf node with a depth limit d_{max} we obtain a sequence of bipartite directed graphs, each consisting of the edges from level d to level $d+1$ of G up to d_{max}. Edges that connect vertices within the same level are not included in the output, but the number of shared links c_i^{sh} of a node i with any other node of the sub-graph, including links between nodes of the same level, are counted during in the BFS. Through tuning of the depth limit d_{max}, it is possible to determine the granularity of the categorization.

A limitation of this approach is that the number of categories found in a BFS search grows exponentially with the depth limit d_{max} due to the positive average node degree of the category graph. To mitigate this limitation, the k most frequent category nodes of each level are chosen based on their frequency c_i^{sh}.

Finally, a number of heuristic patterns were used to extract more generic topics from composite WIKIPEDIA categories such as *Artists_of_Modernism_by_Country* with *Artist* and *Modernism* being the high level topics to be extracted. Table 2 shows a subset of regular expressions which were recursively applied to decompose composite categories. Categories for which no pattern matches are split into words and common stop-words are removed.

Through experimentation, we used $n = 4$ granularity levels with their $k = 5$ most frequent categories, resulting in 652, 235, 104, and 53 thousand categories in level 1, 2, 3, and 4 respectively. A total of 44 patterns and 321 stop-words were used. Table 3 shows the mapped categories for the English *Covid-19* article.

Pattern
`^(\w +)_and_the_(\w +)$`
`^(\w +)_and_(\w +)$`
`^(\w +)_of_the_(\w +)_by_country$`
`^(\w +)_of_(\w +)_by_country$`
`^(\w +)_of_the_(\w +)$`
`^(\w +)_of_(\w +)$`

Table 2: Subset of regular expressions to recursively extract high level categories from WIKIPEDIA compound categories.

4.5 Data Product Visualization

[11]hidden category nodes are nodes that share an edge with the hidden node with page ID 15961454

Level 1	Level 2	Level 3	Level 4
Covid-19	Health policy	Animal health	Social policy
Occupational safety	Airborne diseases	Humanities	Medicine
Health	Infectious diseases	Social sciences	Epidemiology
Zoonoses	Safety	Public policy	Natural sciences
Viral respiratory tract infections	Mode	Diseases	Anthropology

Table 3: Categorization for the English *Covid-19* article with page ID 63030231 for granularity's 1 through 4.

An interactive, web-based visualization[12] of the data product obtained in this study was designed to facilitate the exploration of the indirect impact of COVID-19 through WIKIPEDIA page views.

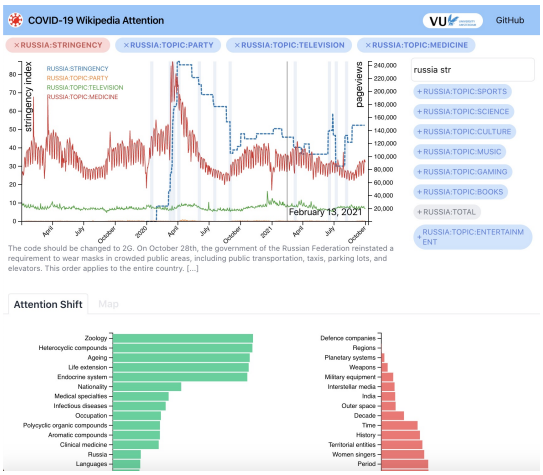


Figure 7: Interactive Data Product visualization.

A screenshot of the visualization is shown in Figure 7. The main component of the visualization is an interactive timeline, in which the user can choose to plot

- The stringency index (SI) of a country with change-points highlighted as vertical bars
- The total WIKIPEDIA traffic for a country based on the primary language in that country
- Page views for selected topics in a country based on the primary language in that country

Of the latter, the user is free to combine as many as desired and compare e.g. stringency with total page views, stringency in different countries, topic page-views for a topic in different countries, or many more. The selection is done by means of tags, which can be filtered based on a text search field.

[12]`https://romnn.github.io/lsde2021/`

If just one stringency index is plotted, hovering over the timeline displays summaries of the most important events with respect to COVID-19 measures. If the user wishes to read more for a specific date, the following cursor can be locked by a single click and unlocked by a subsequent click.

At any time, the user can click on a change-point to show the 10 topics with the highest positive and negative shift in attention for the selected change-point in two bar plots under the main timeline.

4.6 Results

4.6.1 Research Questions

- **RQ1** Does COVID-19 and its measures have an impact on information seeking patterns? (what information is looked up how much?)
- **RQ2** Are those trends subject to temporal or spatial differences?

Our hypothesis for RQ1 was two-fold, we expected to observe a shift in attention to virus related pages at the beginning of COVID-19 and we expected to see a shift in attention depending on the level of the stringency index. The first, we do observe in the results (see plot viruses). The latter, however we do not observe in the data. We reason that the difference in result has to do with the scale and duration of impact. Whereas at the beginning of COVID-19 many people were potentially scared and experienced panic, the pandemic started to become the new normality and people adapted. Therefore the possible impact of stringency change-points was less visible.

Following RQ1, we believed for RQ2 that the trends would show temporal and spatial differences. Based on our results of the difference-in-difference regression we can reject this hypothesis. Because we did not observe significant impact during the other COVID-19 change-points and therefore we conclude that they follow the same spatial and temporal trend [13].

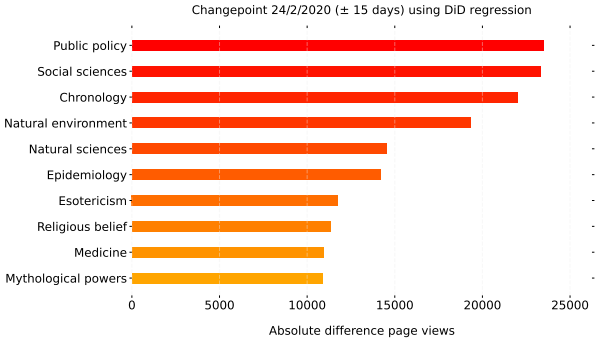


Figure 8: Topics with largest increase in absolute attention based on difference-in-difference regression for the change-point on February 24th 2020 in Italy.

The change-point on February 24th 2020 in Italy shown in Figure 8 shows a trend that we could often observe for larger countries. Especially during the beginning of the pandemic, where not much was known about the virus and policy makers reacted with very strict measures, information seeking was either highly targeted towards practical information (e.g. *Public policy*, *Epidemiology*, or *Medicine*) as well as the emotional sorrows and fears of people (e.g. *Mythological powers*, *Esotericism*, or *Religious belief*), or both, depending on the country.

Looking at the total number of traffic for the Italian Wikipedia with respect to the stringency index in Italy as shown in Figure 11, we also note that during the beginning of the

[13] see section 4.8 for the limitations

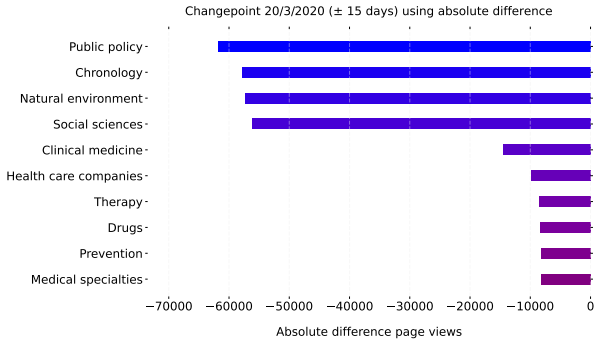


Figure 9: Topics with largest decrease in absolute attention based on absolute difference for the change-point on March 20th 2020 in Italy.

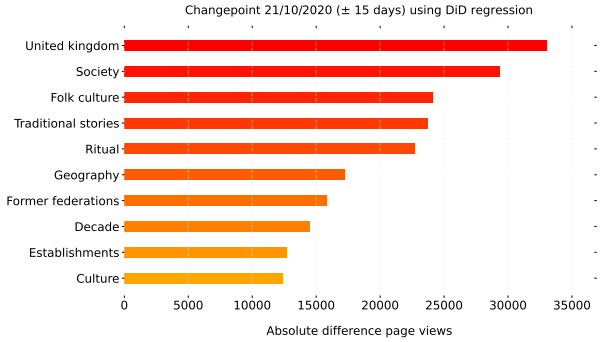


Figure 10: Topics with largest increase in absolute attention based on difference-in-difference regression for the change-point on October 21th 2020 in Italy.

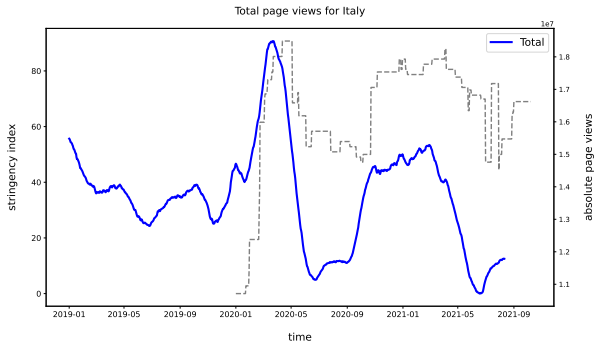


Figure 11: Total traffic for the Italian Wikipedia and the Italian stringency index from 2019 to 2021.

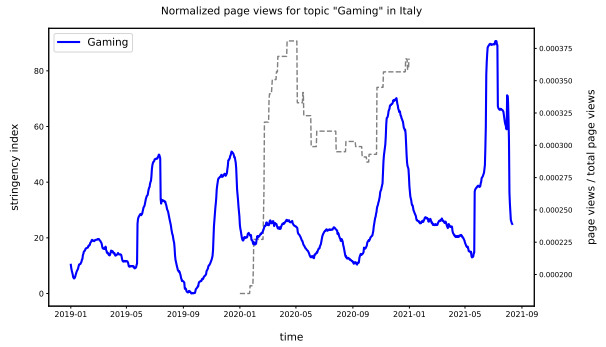


Figure 12: Normalized page-views for the topic "Gaming" for Italy and the Italian stringency index from 2019 to 2021.

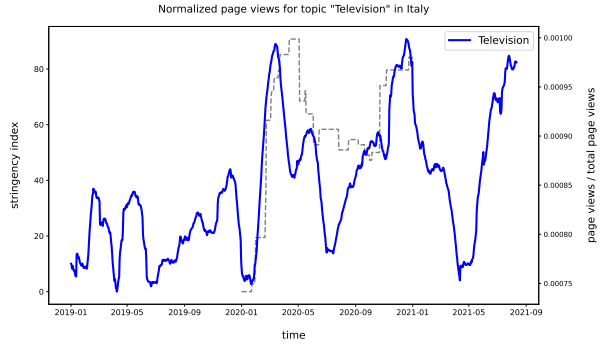


Figure 13: Normalized page-views for the topic "Television" for Italy and the Italian stringency index from 2019 to 2021.

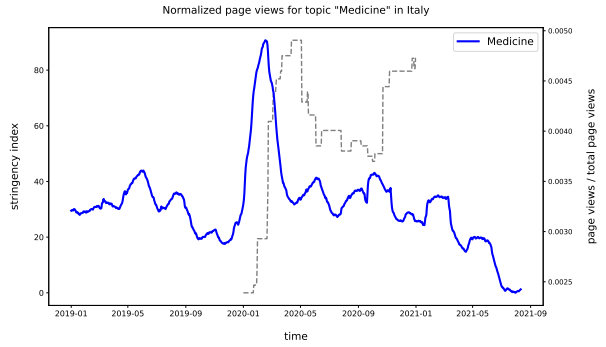


Figure 14: Normalized page-views for the topic "Medicine" for Italy and the Italian stringency index from 2019 to 2021.

pandemic, there was a large peak in traffic with COVID-19 related topics accounting for the most absolute difference. However, as the pandemic continued, conditions quickly went back to normality. The inverse trend can be observed when looking at the topics with the largest absolute decrease in attention for the change-point on March 20th 2020 about a month later as shown in Figure 9. Even when the stringency peaks a second time in Italy on October 21st 2020 as shown in Figure 10, other topics such as *United Kingdom*, most likely due to the Brexit[1] deal, are most relevant.

Considering normalized total page-views for specific topics, we were able to reproduce some findings from Ribeiro et. al (2020)[9] introduced in Section 2.2.2. In Italy, a growing increase in attention for the entertainment related topics *Gaming* and *Television* can be observed as depicted in Figure 12 and **??**, a short spike in attention for the topic *Medicine* as shown in Figure 14, however, no decrease for the topic *Sports*.

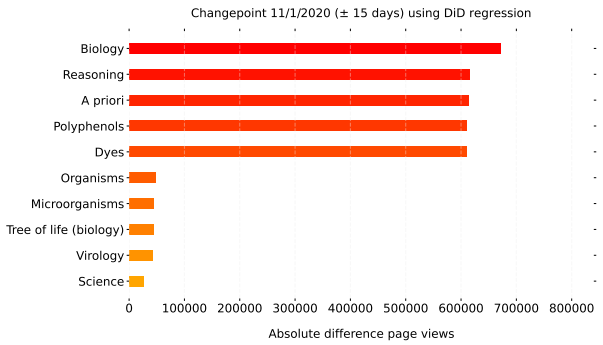


Figure 15: Topics with largest increase in absolute attention based on difference-in-difference regression for the change-point of the first death in China on January 11th 2020 in Germany.

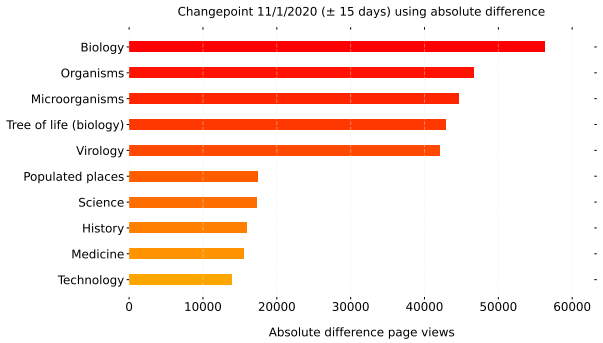


Figure 16: Topics with largest increase in absolute attention based on absolute difference for the change-point of the first death in China on January 11th 2020 in Germany.

It must be noted that the few examples presented reflect general trends we observed for many countries, but we do not claim their generality. Apart from the beginning months of the pandemic, no obvious trends with respect to COVID-19 (stringency) could be observed, especially for topics with the largest decrease in attention. For very small countries, there were hardly any trends, which we attribute to the low number of page-views and the associated noise. For further exploration of the data, the reader is encouraged to use the interactive visualization.

When comparing the topics with largest increase in attention based on difference-in-difference (Figure 15) and absolute difference (Figure 16) for the change-point of the first death in China on January 11th 2020, we observe that in some cases, the absolute difference and difference-in-difference approaches result in slightly different rankings. While it is generally difficult to say which one is more accurate, we assume that absolute difference is more adaptable to rapid changes in topics.

4.7 New Questions

When interpreting the (absence of) results of shifts in attention during the COVID-19 pandemic, we ask the following new question: Is a stringency index change-point a good indicator for capturing shifts in local attention shifts. It could be that the level of international media coverage (social media, American news) or mobility indexes are stronger indicators to capture attention shifts. Furthermore, due to the increasing globalization we have asked ourselves if our second research question is even still relevant today. Is it even possible to capture the effect of any type of event in per country attention shifts based on WIKIPEDIA data?

4.8 Limitations

In our results we could not observe the impact of stringency change-points on attention shifts. Although this could be the conclusion, the reader has to keep in mind the following limitations while interpreting the results.

- **Difference-in-difference** Although difference-in-difference regression can be powerful, it assumes a linear trend. When comparing 2019 (control) to 2020 (treatment) some yearly trends might be similar, yet the trends will not be perfectly linear. Further studies could explore more advanced data-science techniques.

  Because a quantitative comparison of difference-in-difference and absolute difference is difficult due to the subjective categorization and subjective interpretation of results, we conclude that based on qualitative assessment both methods resulted in similar rankings of attention shifts, while the absolute difference approach included more temporally relevant topics.

  Generally, we found that considering relative differences in page views emphasized outliers with very low page views too much and were not representative for changes in topic attention of the mass population.

- **Categorization** Considering our research question, good categorization constitutes the foundation for insightful analysis of the raw page-view data. Despite the subjectively good results obtained by the algorithmic category mapping proposed in this study on many samples, the number of mapped categories is still very large even for granularity level 4. Furthermore, important categories frequently are mapped to lower granularity and are not considered. Perhaps this introduced noise that made it difficult to isolate possible effects. Future studies could reiterate the study with a different category mapping, such as the ORES `articletopic` prediction model. This could help examine the trade-off between topic count, noise, and too generic topics.

5. CONCLUSIONS

Taking in consideration the above mentioned limitations, we conclude that only the start of the pandemic did have a significant impact on the shift in attention as hypothesized. The goal of this study was to complete the picture on the indirect impact of COVID-19 and its measurements. To the best of our knowledge, the stringency index was a neglectable factor in causing attention-shifts most of the time.

We learned that it is difficult to make sense of raw page-view data, as results are hardly verifiable and attention on the internet is subject to many inter-correlated influences which would warrant historical context of all important events on a local and global scope.

One interesting aspect for policy makers could be that the long-duration of the pandemic does not lead to massive attention shifts (shock-effects) over and over again. Naturally, one should keep in mind that the absence of for example social activities for a longer period of time could lead to other negative (health) impacts.

References

Brexit deal. https://www.reuters.com/article/uk-britain-eu-barnier-idUKKBN25M26W?edition-redirect=uk.

ORES articletopic. https://www.mediawiki.org/wiki/ORES/Articletopic.

ORES mediawiki. https://www.mediawiki.org/wiki/ORES.

Ruptures change point detection. https://centre-borelli.github.io/ruptures-docs/. Accessed: 2021-10-01.

K. Abay and K. Tafere. Winners and losers from covid-19 : Global evidence from google search. 2020.

B. et al. Economic and social consequences of human mobility restrictions under covid-19. *Proceedings of the National Academy of Sciences*, 2020.

T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, et al. A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature Human Behaviour*, 5(4):529–538, 2021.

A. Panisson, D. Paolotti, and C. Cattuto. The impact of news exposure on collective attention in the united states during the 2016 zika epidemic. 2020.

M. Ribeiro, K. Gliroric, M. Perard, F. Lemmerich, M. Strohmaier, and R. West. Sudden attention shifts on wikipedia during the covid-19 crisis. 2020.

B. J. Ryan, D. Coppola, D. V. Canyon, M. Brickhouse, and R. Swienton. Covid-19 community stabilization and sustainability framework. *Disaster Medicine and Public Health Preparedness*, 2020.

A. Spark. Apache spark. *Retrieved January*, 17:2018, 2018.

T. N. Y. Times. Coronavirus timeline, 2021. Accessed: 2021-10-01.

R. S. Xin, J. E. Gonzalez, M. J. Franklin, and I. Stoica. Graphx: A resilient distributed graph system on spark. In *First international workshop on graph data management experiences and systems*, pages 1–6, 2013.