# Sudden Attention Shifts on Wikipedia Following COVID-19 Mobility Restrictions

**Preprint** · May 2020

CITATIONS
0

READS
50

**6 authors**, including:

Manoel Ribeiro
École Polytechnique Fédérale de Lausanne
**27** PUBLICATIONS **192** CITATIONS
SEE PROFILE

Kristina Gligoric
École Polytechnique Fédérale de Lausanne
**11** PUBLICATIONS **21** CITATIONS
SEE PROFILE

Maxime Peyrard
Technische Universität Darmstadt
**19** PUBLICATIONS **216** CITATIONS
SEE PROFILE

# Sudden Attention Shifts on Wikipedia Following COVID-19 Mobility Restrictions

Manoel Horta Ribeiro,[♠,*] Kristina Gligorić,[♠,*] Maxime Peyrard,[♠,*]
Florian Lemmerich,[♣] Markus Strohmaier,[♣,♢] Robert West[♠]
♠ EPFL, ♣ RWTH Aachen, ♢ GESIS
{manoel.hortaribeiro,kristina.gligoric,maxime.peyrard,robert.west}@epfl.ch
{florian.lemmerich,markus.strohmaier}@cssh.rwth-aachen.de



## Abstract

We study how the coronavirus disease 2019 (COVID-19) pandemic, alongside the severe mobility restrictions that ensued, has impacted information access on Wikipedia, the world's largest online encyclopedia. A longitudinal analysis that combines pageview statistics for 12 Wikipedia language editions with mobility reports published by Apple and Google reveals a massive increase in access volume, accompanied by a stark shift in topical interests. Health- and entertainment-related topics are found to have gained, and sports- and transportation-related topics, to have lost attention. Interestingly, while the interest in health-related topics was transient, that in entertainment topics is lingering and even increasing. These changes began at the time when mobility was restricted and are most pronounced for language editions associated with countries, in which the most severe mobility restrictions were implemented, indicating that the interest shift might be caused by people's spending more time at home. Our results highlight the utility of Wikipedia for studying reactions to the pandemic across the globe, and illustrate how the disease is rippling through society.

## 1 Introduction

The coronavirus disease 2019 (COVID-19) pandemic has led to the implementation of unprecedented non-pharmaceutical interventions ranging from case isolation to national lockdowns [1]. These interventions have created massive shifts in people's lives. For instance, at the time of writing, more than a third of the global population is under lockdown [2], and millions have lost their jobs or have moved to work-from-home arrangements [3].

Unlike most previous events that directly impacted so many lives around the world, the COVID-19 pandemic developed in a time of widespread access to the Internet. This digitization enables researchers to measure the impact of the pandemic across society in novel ways, by analyzing how it has impacted users' digital traces. For instance, Wikipedia, the world's largest encyclopedia and one of the most visited sites on the Web, captures rich digital traces from readers and makes them publicly available in aggregated form. Wikipedia is used hundreds of millions of times each day to address a wide spectrum of information needs, ranging from reading up on something that was discussed in the media to getting useful information in order to make personal decisions [4]. Given its widespread and varied usage, as well as the public availability of much of its data, Wikipedia thus has the potential to serve as a sensor for the interests, needs, and concerns of entire societies through time at an unprecedented scale.

*These authors contributed equally.

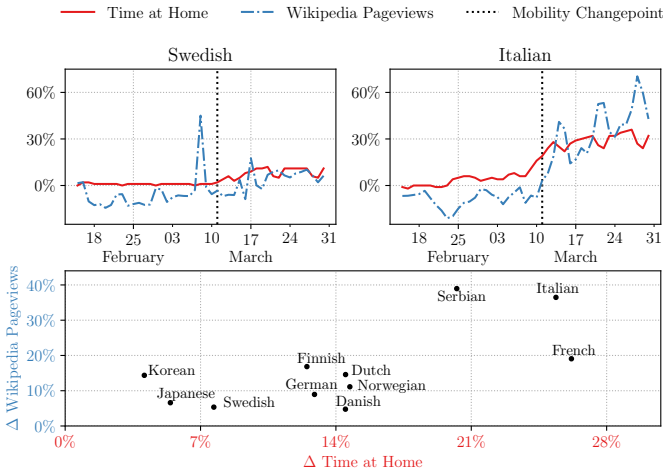


**Figure 1: Wikipedia access vs. mobility.** Association between increase in time spent at home (from Google mobility reports; red) and increase in Wikipedia access volume (from Wikipedia pageview statistics; blue), both in terms of relative change over a 5-week baseline period in early 2020. **Top:** time series for Sweden/Swedish Wikipedia and Italy/Italian Wikipedia; dotted vertical lines: changepoints in mobility time series. **Bottom:** summary for 11 of the 12 languages studied (exc. English); *x*-axis: post-minus-pre-changepoint difference in relative mobility change; *y*-axis: post-minus-pre-changepoint difference in total Wikipedia access volume (Pearson's $r = 0.63$, $p = 0.03$).

**Objective.** In this spirit, the present work aims to elucidate how information access patterns across 12 Wikipedia languages editions have shifted during the current COVID-19 pandemic, in particular in response to the massive mobility restrictions imposed by governments worldwide, as captured in the mobility reports made available by Apple and Google.

**Introductory example.** Consider Fig. 1 (top), which shows how increases in the time spent at home are tracked by increases in Wikipedia usage, for two countries with vastly different reactions to the COVID-19 pandemic: In Sweden, where no lockdown was implemented and mobility decreased only moderately, Wikipedia usage also grew only slightly. In Italy, on the contrary, where a severe lockdown led to pervasive home confinement, Wikipedia usage increased by over 60% compared to a pre-pandemic baseline. The bottom panel of Fig. 1 summarizes the situation across language editions, showing that the effects are not specific to Sweden and Italy.

**Approach and key findings.** Our longitudinal analyses combine Wikipedia access logs with mobility reports made available by Apple and Google (Sec. 2). Methodologically, drawing meaningful conclusions from the longitudinal Wikipedia access logs is challenging due to the presence of trends and seasonalities. We overcome these hurdles by performing careful difference-in-differences analyses in a regression framework. We proceed in three steps:

1. We begin by analyzing the variation of Wikipedia pageview volumes from the early stages of the pandemic until the time of writing. We observe massive increases for all but a few language editions at precisely the time when mobility is restricted (in some cases amounting to a doubling of the number of pageviews), both for COVID-19-related and for other pages (Sec. 3.1).

2. Next, we investigate whether this increase was proportional to the pre-pandemic interest distribution ("more of the same") or whether topics of interest have diverged from normality during the pandemic. We find overwhelming evidence for massive a shift from normality (Sec. 3.2).

3. We further characterize the nature of the attention shift by examining which topics gained and lost attention when mobility decreased, finding that health- and entertainment-related topics won most, whereas sports- and transportation-related topics lost most. We refine the topical analysis further via spatial and temporal analyses (Sec. 3.3).

**Related work.** Our work extends a rich literature on leveraging Wikipedia as a sensor for measuring the behavior of populations and individuals in response to unexpected events, crises, and catastrophes [5–7]. In work that is most closely related, researchers have used Wikipedia pageviews in order to monitor and forecast diseases at a global scale [8, 9] and to study anxiety and information seeking about infectious diseases, such as influenza [10], H1N1 [11], and Zika [12]. COVID-19 is already shadowing these previous epidemics in terms of global damage. Our results shed light on people's shifting interests as they are concerned about the disease and confined to their homes, and they highlight the integral role played by Wikipedia in times of crisis.

## 2 Data

The analyses of this paper combine information about the content and usage of Wikipedia with information about the progression of the pandemic.

**Wikipedia.** We selected 12 languages (cf. Fig. 1) that have large Wikipedia editions and are spoken as the predominant language in European countries severely affected by COVID-19. For each language edition, we obtained publicly available pageview statistics,[1] which specify how frequently each article was accessed daily between 1 January 2018 and 20 April 2020, separately for the desktop and mobile versions of the site. Moreover, we labeled each article with one of 57 topics based on an established topic model (listed in Fig. 5 details in Appendix A.1), and as either COVID-19-related or not.[2] The number of COVID-19-related articles varies substantially across different language editions; e.g., for English over 300 different articles are labeled as associated with the disease, whereas for Swedish, there are only 9.

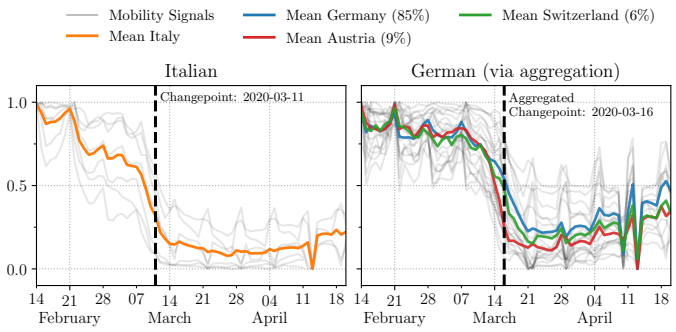


**Figure 2: Robust detection of mobility changepoints from Google and Apple mobility reports.** The reports specify by what percentage the time spent at various location types has changed, compared to a baseline period in January 2020. We use time series for all location types, with multiple changepoint detection algorithms and varying hyperparameters. Given the abruptness of the decrease in most countries, different runs largely agree; we use the average as a robust changepoint. For languages spoken in multiple languages, e.g. German (right), we average mobility time series from the main countries where the language is spoken (here 3), weighted according to the proportion of native speakers (details in Appendix A.2).

**Pandemic timeline.** Nine of the 12 languages are primarily spoken in a single country. For each country, we used Wikipedia to manually determine 5 days of particular interest in the context of COVID-19: first reported case, first death, ban of public events, school closure, lockdown.

Two problems with employing the aforementioned days of interest in statistical analyses are (1) that it is not guaranteed that they would impact movement patterns across different countries homogeneously (e.g., it could be that for some of the countries people stayed more at home even before the lockdown was enacted), and (2) that not all days of interest were observed for the countries of interest (e.g., in Sweden, Korea, and Japan, there was no country-wide lockdown between 1 January and 20 April 2020).

**Mobility reports.** We turn to daily mobility reports published by Apple and Google,[3] which capture population-wide movement patterns based on cellphone location signals. The mobility reports specify, for each day, by what percentage the time spent in various location types (e.g., residential areas, workplaces, retail and recreation, etc.) differed from a pre-pandemic baseline period in early 2020. Both government-mandated lockdowns as well as self-motivated social distancing measures manifest themselves as sharp changes in the mobility time series, which we detect automatically using changepoint detection algorithms (Fig. 2, left; details in Appendix A.2). We henceforth refer to these points as *mobility changepoints.* We use mobility changepoints as heuristic

[1] `https://dumps.wikimedia.org/other/pageviews/`

[2] Based on the article list of `https://covid-data.wmflabs.org`.

[3] `https://www.(apple|google).com/covid19/mobility/`

dates for when people started spending substantially more time in their homes. Unlike choosing one of the days of interest, this leads to a meaningful "treatment" across different countries.

Three of the 12 languages are spoken more widely than in one predominant country: English, German, and French. As Wikipedia pageview statistics are available only at the language level, not at the country level, we determine a mobility changepoint for these language editions by aggregating mobility reports for the countries in which the language is official (Fig. 2, right; details in Appendix A.2).

We emphasize that, since any Wikipedia edition can be accessed from anywhere (barring censorship), the link between Wikipedia language editions and countries is merely approximate, even for languages that are official in only a single country. This should be kept in mind when interpreting our results, especially for the English edition, which is read widely across the globe.

# 3 Results

## 3.1 Shifts in overall pageview volume

**COVID-19-related pageviews.** Wikipedia has been shown to be an accurate and up-to-date source of COVID-19-related information,[4] and we start by investigating how heavily this information was accessed by users. Fig. 3 (column 1) tracks the popularity of COVID-19-related articles (as a fraction of all pageviews) in all 12 languages over the course of the pandemic, from 14 January (the earliest time for which mobility reports are available) to 20 April 2020. In nearly all languages, the share of COVID-19-related pageviews increased up to the mobility changepoint (dashed vertical line), from where on it tended to decrease.[5] We emphasize that these time series are plotted on logarithmic $y$-scales, such that a linear slope in the plots, even if small, corresponds to an exponential rate of change.

COVID-19-related articles were generally among the most popular during the period of study; e.g., 12 of the 15 most accessed articles in the English desktop version were related to the pandemic. In some languages, the fraction of pageviews going to COVID-19-related articles surpassed 2% on some days, a considerable share of Wikipedia's overall volume, considering that Wikipedia has over 6 million articles.

**Non-COVID-19-related pageviews.** Next, we focus on Wikipedia articles not related to COVID-19, the vast majority. Their popularity during the period of study, in terms of the daily total number of pageviews, is shown as solid lines in Fig. 3 (column 2; linear $y$-scales). The dotted lines correspond to the same period in 2019, precisely one year earlier.[6] Access volumes in 2020 closely mirror those in 2019 up to the mobility changepoint associated with the respective language. Thereafter, the access volumes of 2020 begin to rise up and above those for 2019 for nearly all languages. The trend is particularly strong for languages spoken in the countries with the most severe lockdown measures (Serbian, Italian, French), and it is weakest for the languages of Japan and Scandinavia (Danish, Swedish, Norwegian, Finnish), where lockdown measures have been weaker than elsewhere.

The difference in trends between countries with a stricter lockdown and those with a weaker or no lockdown can also be seen in Fig. 1. There, in the top two plots we can see the relative change in Wikipedia pageviews and time spent at home for Swedish and Italian. Comparing the two, we find that the increase of both time at home and Wikipedia pageviews was larger for Italian. In the bottom half of the figure, we plot the increase (after vs. before the mobility changepoint) in time spent at home ($x$-axis) against the increase in total Wikipedia pageview volume ($y$-axis) for all 12 languages, observing that the two values are significantly correlated ($r = 0.63$, $p = 0.03$) and witnessing again the difference between countries with stricter lockdown measures (Serbia, Italy, France) and those with weaker or no lockdown measures (Korea, Japan, Sweden). The positive correlation between time at home and pageviews suggests that lockdowns have impacted Wikipedia reading habits.

We further explore this relationship in Fig. 3 (column 3) via cumulative plots, where the daily 2020-minus-2019 difference is accumulated from 14 January onward. Since languages have vastly different access volumes overall (cf. column 2), and in order to have a common scale across languages, we express the cumulative difference in terms of multiples of the average daily pageview volume attained in 2019 by the respective language. The increase is dramatic for some languages; e.g., the Serbian Wikipedia edition has experienced about 30 days' worth of surplus access volume between mid-March and mid-April 2020, corresponding roughly to a doubling in access volume, compared to the previous year. Strong effects are also observed for Italian (15 surplus days), French (10 days), English (6 days), Dutch (6 days), and Korean (5 days).

Note that 8 of the 12 languages initially ran a deficit, with Wikipedia being visited less in 2020 compared to the corresponding days in 2019 (reflected as negative values in the plots of column 3 of Fig. 3), but all except 3 languages (Norwegian, Swedish, Danish) recovered and eventually ran a surplus by the end of the study period.

**Difference-in-differences regression.** In order to go beyond visual inspection and to quantify the differences between languages more objectively, we take a regression-based difference-in-differences approach [13].

In this setup, we consider, for each language a time window of 10 weeks (70 days) centered around the respective mobility changepoint in 2020, as well as the corresponding time window in 2019. Each of these 140 days contributes one data point per language, for a total of $140 \times 12 = 1{,}680$ data points. As the dependent variable $y$, we use the logarithm of the number of pageviews, and as independent variables, the following three factors: `year` (2019 or 2020), `period` (before or after calendar day of mobility changepoint), `language`. We now model $y$ as a linear function of these three factors and all their two- and

[4] https://wikimediafoundation.org/covid19/

[5] Some languages saw extreme upticks on certain days. Most of them can be linked to the creation of important COVID-19-related articles; e.g., the Swedish article CORONAVIRUSUTBROTTET 2020 I SVERIGE (English: COVID-19 PANDEMIC IN SWEDEN) was created on 22 March 2020 and immediately received wide attention.

[6] Since Wikipedia access volumes tend to follow a weekly periodicity, our alignment was manually adjusted to ascertain that matched days correspond to the same day of the week.

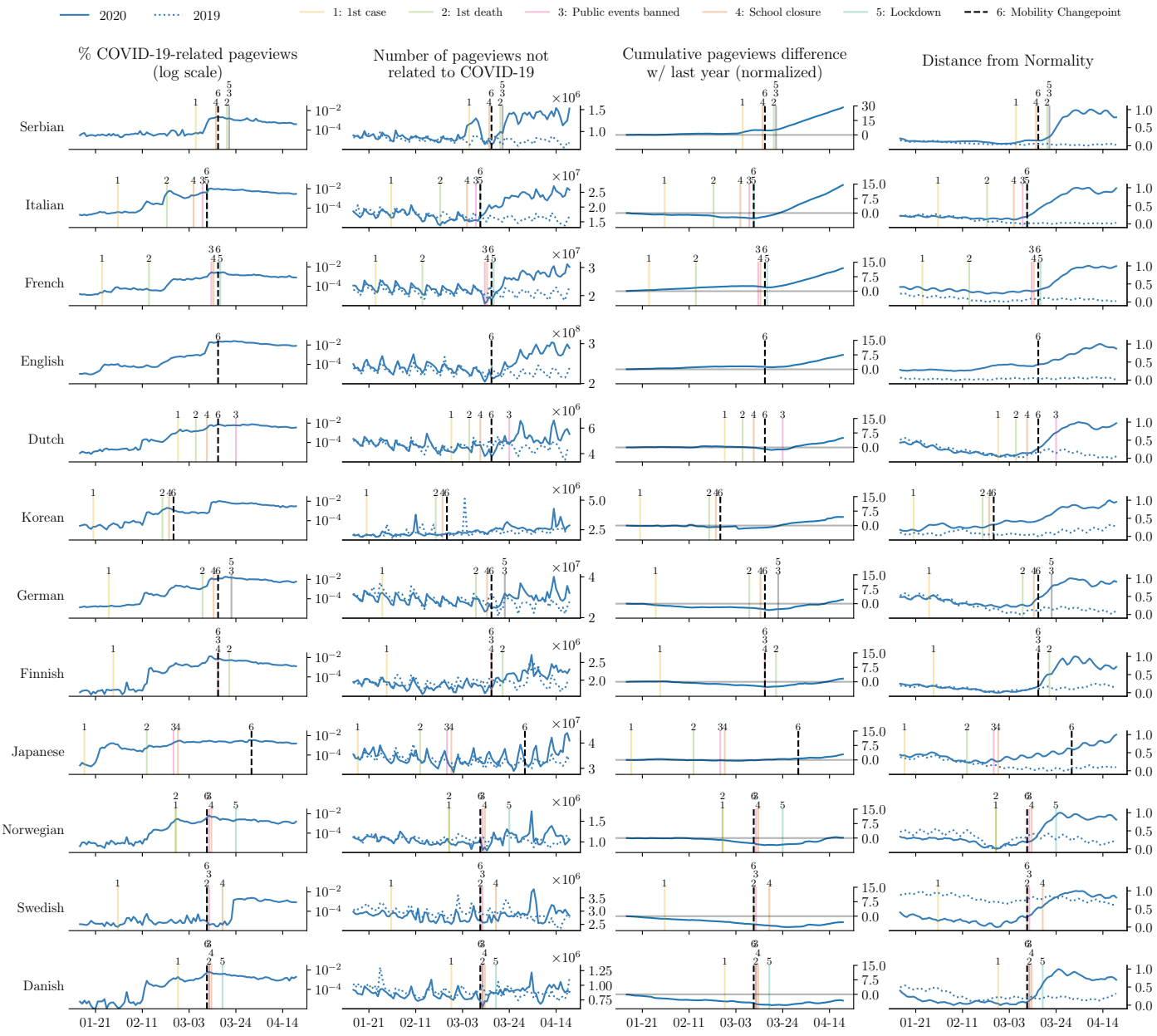


**Figure 3: Evolution of Wikipedia access for 12 language editions from mid-January to mid-April 2020. Column 1:** percentage of pageviews spent on COVID-19-related pages (logarithmic scale). **Column 2:** total number of pageviews for all other pages (linear scales); solid: 14 January to 20 April 2020; dotted: corresponding period in 2019. **Column 3:** cumulative difference in total number of pageviews between study period of 2020 and corresponding period of 2019, reported in terms of multiples of average daily number of pageviews in 2019 ($y$-axes identical across languages, except for Serbian, where cumulative volume is twice as high). **Column 4:** "distance from normality" with respect to topical attention (cf. Sec. 3.2) for 2020 (solid) and 2019 (dotted). All plots show, as vertical lines, mobility changepoints and pandemic-related events. For languages spoken in multiple countries, mobility changepoints are aggregated as described in Fig. 1, and events are shown for the country with the largest number of speakers (omitted for the globally popular English edition).

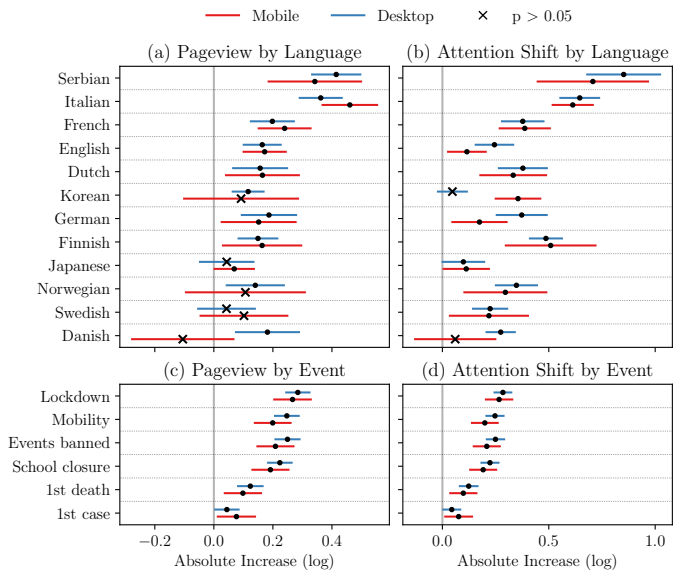


**Figure 4: Estimated effects of restricted mobility.** Results of difference-in-differences regressions for estimating effects of restricted mobility on **(a)** total number of pageviews and **(b)** topical attention, for 12 studied language editions are shown. **(c–d)** Effects are pooled over all languages, but computed for different cutoff events as "treatments" (panels (a) and (b) correspond to "mobility"). Error bars: 95% CIs approximated as 2 standard errors. Dependent variables are used in logarithmic form, such that exponentiated coefficients capture the multiplicative increase due to the treatment. Recall that, since $x \approx e^x - 1$ for $x \approx 0$, the logarithmic effects roughly approximate the percentage increase.

three-way interactions. In R formula notation,

$$y \sim \texttt{year} * \texttt{period} * \texttt{language}. \quad (1)$$

Here, $\texttt{a} * \texttt{b}$ is shorthand notation for $\texttt{a} + \texttt{b} + \texttt{a} : \texttt{b}$, where in turn $\texttt{a} : \texttt{b}$ stands for the interaction of a and b.

Pageview volumes were considered in logarithmic form for two reasons: first, because raw pageview counts are far from normally distributed, with numerous large outliers, and second, because the logarithm makes the model multiplicative, thus implicitly normalizing the estimated effects and making it possible to compare languages with vastly different absolute pageview volumes: if $b$ is the coefficient of the three-way interaction $\texttt{year} : \texttt{period} : \texttt{language}$, then $e^b$ captures the multiplicative factor by which pageview volumes increased when mobility dropped, after accounting for differences stemming from the year alone or the period alone,[7] which are already captured by the coefficients of $\texttt{year} : \texttt{language}$ and $\texttt{period} : \texttt{language}$, respectively.

The estimated logarithmic effects are plotted for all languages in Fig. 4(a), separately for the mobile vs. desktop versions of Wikipedia. The results confirm the effects observed visually in Fig. 3 (column 3); e.g., the logarithmic pre-vs.-post lockdown effect on the Italian desktop version is around 0.48 (corresponding to an increase in pageviews to $e^{0.48} \approx 162\%$) whereas the effects are smaller and mostly insignificant for Danish, Swedish, Norwegian, and Japanese.

Additionally, we fit a slightly different model:

$$y \sim \texttt{year} * \texttt{period} + \texttt{language}, \quad (2)$$

where the dependent variable $y$ is again the logarithm of the number of pageviews, but language now merely serves as a language-specific baseline, and $\texttt{year} : \texttt{period}$ is the only interaction term. This way, the coefficient of $\texttt{year} : \texttt{period}$ captures the estimated effect in a language-independent way. We fit this model not only for the case where period is defined by the mobility changepoint, but also where it is defined by the 5 other pandemic-related events (first case, first death, etc.). This way, we may compare effect sizes for the various events when used as "treatments".

The estimated logarithmic effects are plotted for each event in Fig. 4(c) ($R^2 > 0.95$ for all models). We find that, for both the mobile and desktop versions, events that are more tightly related to decreased mobility are associated with the largest increases in pageviews. For example, for mobile, the pre-vs.-post first-death effect is only around 0.12 ($e^{0.12} \approx 127\%$); for school closure, it is around 0.22 ($e^{0.22} \approx 124\%$); and for the actual lockdown, it is largest, at 0.28 ($e^{0.28} \approx 132\%$).

This corroborates the previous results on mobility and Wikipedia access from yet another angle. Earlier, in Fig. 3 and Fig. 4(a) we showed that languages spoken in countries with a stricter lockdown saw a larger pageview increase, whereas here we showed that, considering all countries, events associated with a mobility decrease are significantly more associated with an increase in pageviews than other pandemic-related events.

## 3.2 Shifts from normality

So far, we observed an overall increase in pageviews. Next, we ask whether this increase boosted pageviews evenly across articles or whether users' attention shifted from certain topics to others.

**Distance from normality.** To quantify attention shifts, we introduce a notion of "distance from normality", as follows. On each day, the pageviews in a given language edition form a distribution over articles, which characterizes how users' attention was distributed on that day. We represent each daily distribution as an "attention vector" with one entry per article and entries summing to 1.

With over 6 million Wikipedia articles, many of which are rarely visited, attention vectors are large and noisy. Therefore, we first apply principal component analysis (PCA) in order to project attention vectors into a low-dimensional subspace. In the subspace, two attention vectors are naturally compared via their Euclidean distance.

The notion of "normal" attention is captured by the average attention vector over all days of 2019, i.e., well before the pandemic; and for each subsequent day, the distance from normality is given by the Euclidean distance of that day's attention vector from the average attention vector.

Distances from normality are plotted as time series in Fig. 3 (column 4). We also plot a baseline time series computed on data from exactly one year earlier (dotted line), where normality is

[7] For instance (cf. Fig. 3, column 2), Swedish pre-mid-March pageviews were consistently lower in 2020 than in 2019; and Finnish post-mid-March pageviews were higher than pre-mid-March pageviews even in 2019, without the pandemic.

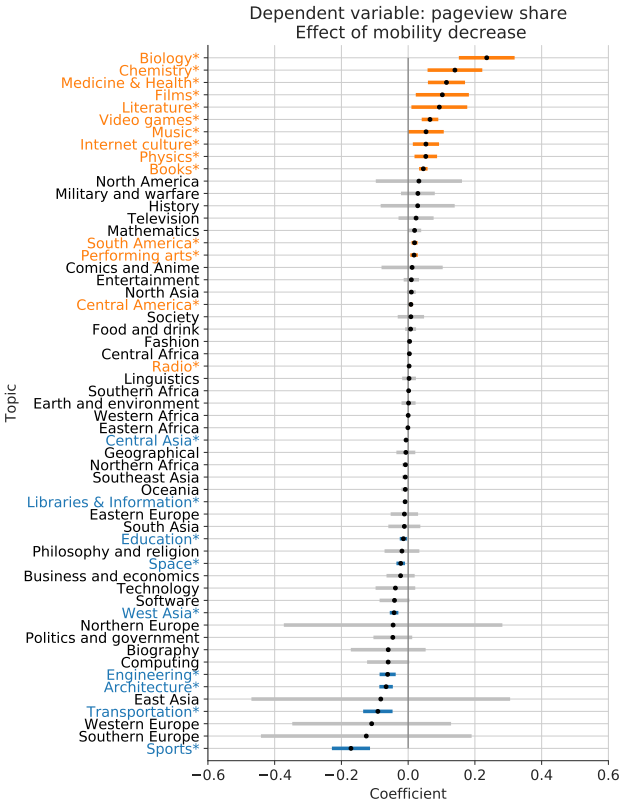


**Figure 5: Topical attention shift.** Effect of mobility decrease on relative pageview share of 57 topics, estimated via difference-in-differences regression, pooled across 12 languages. Error bars: 95% CIs. $*p < 0.05$ (two-sided); significant positive (negative) coefficients in orange (blue).

defined as the average attention vector of 2018, and distance from normality was computed for mid-January to mid-April 2019. We observe that the (dotted) 2019 baseline curves stay flat around the mobility changepoints. If additional pageviews gained after the changepoints mirrored the average attention distribution, one would expect equally flat curves for 2020. Quite on the contrary, however, the curves for 2020 increase sharply at the mobility changepoints, a clear indication of large topical shifts in overall attention.

We also performed the same analysis as in Fig. 1, calculating the $\Delta$(Distance to Normality) instead of $\Delta$(Time at Home). We observe very similar results ($r = 0.52, p = 0.08$), with similar relative positions for all languages: Serbian, Italian, and French on the top right; Korean, Japanese, and Swedish on the bottom left. We omit this figure for space reasons.

**Difference-in-differences regression.** To quantify the attention shifts more objectively, we perform a difference-in-differences regression analogous to the one presented in Sec. 3.1 (Eq. 1), but this time with the outcome $y$ being the logarithm of the distance from normality, rather than of the pageview volume ($R^2 = 0.95$). The three-way interaction terms, visualized in Fig. 4(b), capture the effect sizes for the various language versions. We observe a significant increase for all 12 languages except for mobile Korean and desktop Danish. In several languages, the attention shifts are drastic; e.g., in Serbian, the increase in distance from normality around the mobility changepoint in 2020 is over twice as large as the corresponding increase in 2019 (multiplicative increase: $e^{0.85} \approx 233\%$).

Mirroring the analysis of Fig. 4(c), we consider all languages and use as the treatment different pandemic-related events, as shown in Fig. 4(d) ($R^2 > 0.95$ for all models). We see a similar picture to the analogous analysis for pageviews (Fig. 4(c)): events that are more tightly related to decreased mobility are associated with highest increase in distance from normality; e.g., the pre-vs.-post first-death effect is only 0.08 ($e^{0.08} \approx 108\%$), whereas the effect of the actual lockdown is 0.24 ($e^{0.24} \approx 127\%$). This matches the results from the pageview analysis: as mobility was reduced, there was a change in the total number of pageviews (how much users visit Wikipedia) as well as in the topical access patterns of users (what articles users visit on Wikipedia).

**Sensitivity to design choices.** As a robustness check, we replicated the distance-from-normality computation for various numbers of principal components (a hyperparameter), observing only negligible effects on the final outcome. Additionally, we also measured distance from normality in a different, information-theoretic way, as the Kullback–Leibler divergence between the daily and average attention distributions. The results are consistent with those presented here, so we omit them for brevity's sake.

## 3.3 Shifts in topic-specific pageview volume

So far, we observed a massive increase in pageview volume (Sec. 3.1), which was not explained by a simple proportional increase according to the prior attention distribution. Rather, the increase in pageview volume was accompanied by major shifts in attention (Sec. 3.2). Next, we investigate the shift further, with the goal of identifying which topics gained, and which lost, attention.

Certain topics may have gained attention due to overall interest trends (e.g., interest in soccer-related topics increasing in years with important international championships), or may be subject to annual seasonality (e.g., food-related topics peaking around Christmas). Therefore, one cannot naïvely compare pageviews before vs. after the mobility changepoints. Instead, we account for trends and seasonality by formulating difference-in-differences regression models analogous to that of Eq. 1.

**Overall topical attention shift.** First, we aim to quantify the change in topical interest independent of languages. We hence use a model without a language term, but with an added topic term ($R^2 = 0.66$; for notation, cf. Eq. 1):

$$y \sim \texttt{year} * \texttt{period} * \texttt{topic}. \tag{3}$$

We code the 57 topics as 56 dummy variables, with one arbitrary topic serving as a baseline topic. The outcome variable $y$ of interest is the fraction of pageviews going to articles of the respective topic on the respective day.

Summing the coefficient of `year : period : topic` with the coefficient of `year : period` (corresponding to the baseline topic) reveals changes in topical interest around the mobility changepoints. We observe (Fig. 5) that the topics that are subject to the largest positive effects (BIOLOGY, CHEMISTRY, and MEDICINE & HEALTH) are all related to the pandemic. These topics are

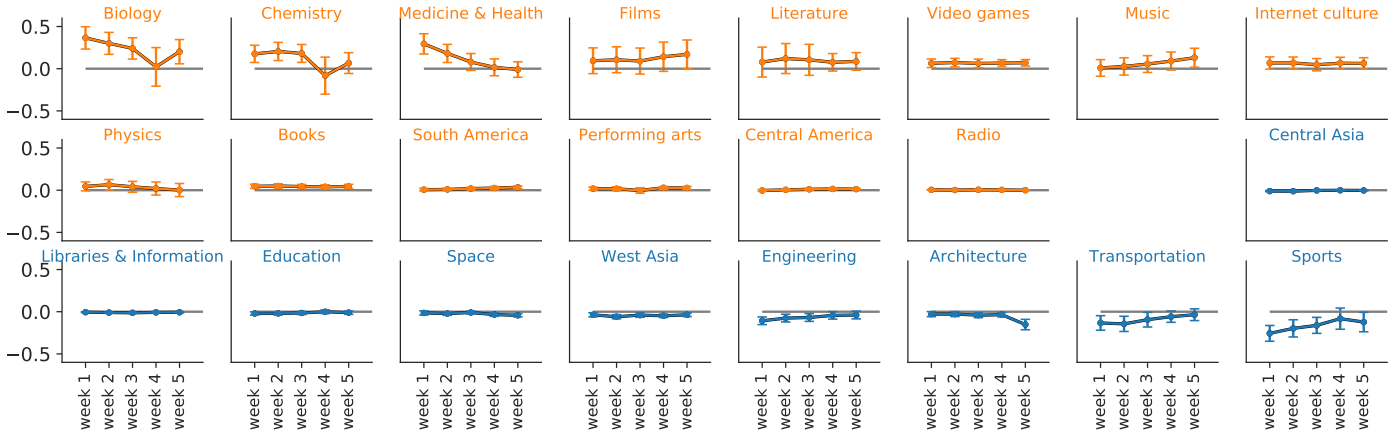


**Figure 6: Temporal evolution of topical attention shift.** Effect of mobility decrease on relative pageview share of 23 topics (with significant coefficients in time-independent model of Fig. 5) in the five weeks following mobility changepoints, estimated via difference-in-differences regression, pooled across 12 languages. Error bars: 95% CIs. Topics with overall positive (negative) coefficients in orange (blue).

followed by entertainment-related categories such as FILMS, LITERATURE, VIDEO GAMES, MUSIC, and INTERNET CULTURE. At the other end of the spectrum, the topic that is subject to the largest negative effect is SPORTS, followed by TRANSPORTATION, ARCHITECTURE, and ENGINEERING.

**Temporal evolution of topical attention shift.** Next we investigate how topical interests evolved over time after the mobility changepoint, via a variation of the time-independent model of Eq. 3, where we now replace the binary `period` indicator (before vs. after calendar day of changepoint) with a 6-level categorical factor `week` whose levels $0, 1, \ldots, 5$ indicate in what week after the calendar day of the changepoint the respective day falls (where level 0 marks all days before the calendar day of the changepoint):

$$y \sim \texttt{year} * \texttt{week} * \texttt{topic}. \tag{4}$$

The coefficients of the model ($R^2 = 0.67$) are plotted as time series in Fig. 6 (showing only topics whose effect is significant with $p < 0.05$ in the time-independent model of Eq. 3; cf. colored topics in Fig. 5). We see that the previously observed time-independent effects are robust, with consistent signs over time. Additionally, some interesting temporal trends emerge: e.g., the topics with the largest positive effects (BIOLOGY, CHEMISTRY, MEDICINE & HEALTH) decrease over time, and those with the largest negative effects (SPORTS, TRANSPORTATION) increase over time, possibly hinting at a gradual return to normality. The opposite trend holds for some entertainment-related topics (FILMS, MUSIC), whose effect sizes are overall high and have increased even further over time.

**Language-specific topical attention shifts.** Finally, we are interested in language-specific shifts in topical attention. We fit an extended version of the language-independent model of Eq. 3, with an additional `language` factor ($R^2 = 0.95$):

$$y \sim \texttt{year} * \texttt{period} * \texttt{language} * \texttt{topic}. \tag{5}$$

For each language, we sort all topics by their 4-way interaction coefficients and display the 3 topics with the largest positive and negative coefficients, respectively, in Table 1 (all $p < 0.05$). Overall, the above-described general patterns also hold for individual languages; e.g., in 6 of the 12 languages, BIOLOGY shows the strongest positive effect, whereas SPORTS is among the bottom 3 topics in 9 of the 12 languages. Notable exceptions include Korean, where SPORTS (mostly soccer-related articles) gained, rather than lost, most, and Italian and French, where we observe the strongest positive effects for entertainment-related topics such as FILMS and LITERATURE, rather than for health-related topics.

# 4 Discussion and Conclusions

In summary, the COVID-19 pandemic has led to exponential spikes in the access of COVID-19-related articles across 12 language editions of Wikipedia. These articles were among the most read articles in all studied language editions, surpassing 2% of the total pageview volume in some cases—a remarkable share, considering that the number of COVID-19-related articles is small (e.g., about 300 out of 6 million English articles), constituting a vanishing fraction of all articles. Interest in these articles was generally highest around the empirically observed mobility changepoints, where users suddenly started spending much more time at home, in most countries due to government-mandated lockdowns.

While access to pandemic-related articles declined after the mobility changepoints, overall Wikipedia usage grew drastically starting at that very time. For instance, during the month after the mobility changepoint, the overall pageview volume grew to about 200% of the pre-pandemic level for the Serbian edition, to about 150% for Italian and French, and to about 120% for English.

The surplus pageviews do not simply distribute proportionally over pages according to the pre-pandemic attention distribution. Rather, a stark shift in attention occurred, starting abruptly at the time of the mobility changepoint. Health- and entertainment-related topics gained attention share, whereas sports- and transportation-related topics lost massively (e.g., SPORTS by nearly 20 percentage points), reflecting the widespread cancellation of sports events and the overall decrease in mobility.

**Table 1: Language-specific topical attention shifts.** Effect of mobility decrease on topics, estimated via difference-in-differences regression for 12 languages. Top and bottom 3 topics shown per language (all with significant regression coefficients, $p < 0.05$).

| Language | Top topics | Bottom topics |
|---|---|---|
| English | NORTH AMERICA, MUSIC, TELEVISION | SPORTS, POLITICS AND GOVERNMENT, TECHNOLOGY |
| French | FILMS, LITERATURE, MEDICINE & HEALTH | SPORTS, WESTERN EUROPE, TECHNOLOGY |
| German | BIOLOGY, CHEMISTRY, FILMS | WESTERN EUROPE, SPORTS, COMPUTING |
| Korean | SPORTS, BIOLOGY, POLITICS AND GOVERNMENT | EAST ASIA, COMPUTING, FILMS |
| Japanese | INTERNET CULTURE, VIDEO GAMES, MUSIC | EAST ASIA, SPORTS, TRANSPORTATION |
| Finnish | BIOLOGY, CHEMISTRY, MEDICINE & HEALTH | POLITICS AND GOVERNMENT, SPORTS, NORTHERN EUROPE |
| Dutch | BIOLOGY, NORTH AMERICA, MUSIC | WESTERN EUROPE, SPORTS, ARCHITECTURE |
| Norwegian | BIOLOGY, SPORTS, MEDICINE & HEALTH | BUSINESS AND ECONOMICS, SOFTWARE, TRANSPORTATION |
| Danish | BIOLOGY, MEDICINE & HEALTH, CHEMISTRY | WESTERN EUROPE, ARCHITECTURE, WEST ASIA |
| Swedish | BIOLOGY, MEDICINE & HEALTH, LITERATURE | NORTHERN EUROPE, SPORTS, TRANSPORTATION |
| Serbian | PHYSICS, SOCIETY, BIOLOGY | SOUTHERN EUROPE, SPORTS, SPACE |
| Italian | FILMS, LITERATURE, HISTORY | SPORTS, SOUTHERN EUROPE, TRANSPORTATION |

We speculate that the increase in Wikipedia usage is caused by users' spending more time at home. As our findings are observational, such a causal link cannot be conclusively established. Indeed, lockdown measures and the ensuing decrease in mobility coincided with numerous other life changes inflicted by the pandemic, and any of these concomitant changes—rather than the increase in time spent at home—might in principle be the true cause of the changes in Wikipedia access patterns; e.g., one might argue that an increased concern about health issues might drive people to access more encyclopedic information.

We argue, however, that our results provide numerous circumstantial indications of a causal link between spending time at home and reading more Wikipedia. First, access to COVID-19-related articles and access to other pages are phase-shifted: as soon as people spend more time at home, the former drops, whereas the latter rises sharply (Fig. 3).

Second, while health-related topics have regressed down toward pre-pandemic levels during the 5 weeks following the mobility changepoint, entertainment-related topics have remained at a stable high (LITERATURE, VIDEO GAMES, INTERNET CULTURE, BOOKS) or even increased in popularity (FILMS, MUSIC; Fig. 6). That is, while people's health concerns seem to abate, their interest in typically homebound entertainment activities remains, pointing at the latter as a more likely cause of the overall increased pageview volume.

Third, we observe a dose–response relationship, in the sense that the increase in pageviews is strongest for the countries with the strongest, and weakest for the countries with the weakest, mobility constraints (Fig. 1). Moreover, language editions associated with countries with the weakest drops in mobility (Danish, Swedish, Norwegian) see the biggest increase in pandemic-related topics (BIOLOGY, MEDICINE & HEALTH), whereas language editions associated with the most severely mobility-restricted countries (Italian, French, Serbian) favor other topics, especially entertainment-related ones.

Last, even within individual languages, we observed the largest growth in pageviews and the largest shifts in attention around the points in time when mobility drops sharply, not when the first infections and deaths are reported (Fig. 4d).

A further interesting causal question asks in what activities people engage when forced to spend time at home. Answering this causal question would require drawing conclusions from the particular lockdown that we are observing in the context of COVID-19 to general situations of restricted mobility. One idea would be to treat COVID-19 as an instrumental variable [14], i.e., a haphazard event that systematically nudges people to stay at home, while affecting people's interests only via the lockdown measures. The latter requirement, called "exclusion restriction", is unlikely to hold in general; e.g., as stated above, health interests might be increased directly by the disease, rather than because more time is spent at home. Exclusion restriction might, however, arguably hold for certain other topical domains, and indeed it would be interesting to explore whether COVID-19 can be used as an instrumental variable to estimate, e.g., what books people read, what movies they watch, or what dishes they cook when they are forced to, or—let the glass be half-full—when they are given the chance to, spend more time at home.

Regardless of causal implications, our study certainly lets us conclude that Wikipedia plays a crucial role during extended times of crisis not only for directly crisis-related topics, but also—in fact even more so—beyond.

## Acknowledgements

This work sprouted out of a collaborative effort from EPFL-DLAB members to get a grip on what was happening with Wikipedia amidst the COVID-19 pandemic. We would like to thank Akhil Arora, Alberto García Durán, Germain Zouein, Lars Klein, Liangwei Chen and Valentin Hartmann for their help and feedback. Last but not least, this work would not have been possible without Tiziano Piccardi, who not only participated in the initial sprint, but also helped us greatly with obtaining the data, and by sharing his "encyclopedic" Wiki-knowledge with us.

## References

[1] Seth Flaxman et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Report, Imperial College London, 2020.

[2] Conrad Quilty-Harper and Layal Sam Wong Liverpool, Adam Vaughan. Covid-19 news: Coronavirus restrictions to ease slightly in England, 2020.

[3] Thomas Hale, Sam Webster, Anna Petherick, Toby Phillips, and Beatriz Kira. Oxford covid-19 government response tracker. *https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-government-response-tracker*, 2020.

[4] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. Why We Read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[5] Ruth García-Gavilanes, Anders Mollgaard, Milena Tsvetkova, and Taha Yasseri. The memory remains: Understanding collective memory in the digital age. *Science advances*, 3(4):e1602368, 2017.

[6] Brian Keegan, Darren Gergle, and Noshir Contractor. Hot off the wiki: Dynamics, practices, and structures in wikipedia's coverage of the tōhoku catastrophes. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, page 105–113, New York, NY, USA, 2011. Association for Computing Machinery.

[7] Stefan Geiß, Melanie Leidecker, and Thomas Roessing. The interplay between media-for-monitoring and media-for-searching: How news media trigger searches and edits in wikipedia. *New Media & Society*, 18(11):2740–2759, 2016.

[8] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y Del Valle, and Reid Priedhorsky. Global disease monitoring and forecasting with wikipedia. *PLoS computational biology*, 10(11), 2014.

[9] Kyle S. Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M. Hyman, Alina Deshpande, and Sara Y. Del Valle. Forecasting the 2013–2014 influenza season using wikipedia. *PLOS Computational Biology*, 11(5):1–29, 05 2015.

[10] David J. McIver and John S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLOS Computational Biology*, 10(4):1–8, 04 2014.

[11] Yla R. Tausczik, Kate Faasse, James W. Pennebaker, and Keith J. Petrie. Public anxiety and information seeking following the h1n1 outbreak: Blogs, newspaper articles, and wikipedia visits. *Health Communication*, 27:179 – 185, 2012.

[12] Michele Tizzoni, AndrÃ© Panisson, Daniela Paolotti, and Ciro Cattuto. The impact of news exposure on collective attention in the united states during the 2016 zika epidemic. *PLOS Computational Biology*, 16(3):1–18, 03 2020.

[13] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[14] Joshua D Angrist and Guido W Imbens. Identification and estimation of local average treatment effects. Technical report, National Bureau of Economic Research, 1995.

[15] Ahmet Aktay et al. Google covid-19 community mobility reports: Anonymization process description (version 1.0), 2020.

[16] Samaneh Aminikhanghahi and Diane J Cook. A Survey of Methods for Time Series Change Point Detection. *Knowledge and information systems*, 51(2):339–367, 2017.

[17] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective Review of Offline Change Point Detection Methods. *Signal Processing*, 167:107299, 2020.

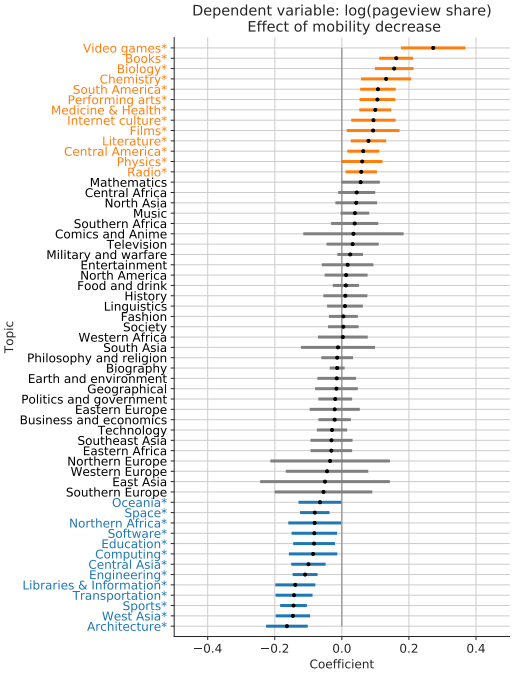


**Figure 7: Overall topical attention shift using multiplicative model.** Effect of mobility decrease on relative pageview share of 57 topics, estimated via difference-in-differences regression, pooled across 12 languages. Error bars: 95% CIs. $*p < 0.05$ (two-sided); significant positive (negative) coefficients in orange (blue).

**Table 2: Changepoint detection results averaged over different parameters choices.** Standard deviations in days.

| Language | Mean | SD | Language | Mean | SD |
|---|---|---|---|---|---|
| English | 03/16 | 0.14 | Dutch | 03/16 | 0.97 |
| French | 03/16 | 0.00 | Norwegian | 03/11 | 0.93 |
| German | 03/16 | 0.07 | Danish | 03/11 | 0.06 |
| Korean | 02/25 | 1.01 | Swedish | 03/11 | 1.02 |
| Japanese | 03/31 | 1.07 | Serbian | 03/16 | 0.89 |
| Finnish | 03/16 | 0.14 | Italian | 03/11 | 0.00 |

# A Appendix

## A.1 Obtaining Topics

We obtain topics for Wikipedia articles using the ORES `articletopic` model.[8] For each page, the top-scoring topic is used as the topical label.

## A.2 Obtaining Changepoint from mobility data

**Mobility Data.** To estimate the effective lockdown dates, we use the mobility reports made available by Google and Apple. Google released community-level reports [15] [9] indicating the daily percentage change in visits to predefined categories of places: *Retail and Recreation* aggregates places like restaurants, cafés, shopping centers, *Grocery and Pharmacy*, *Parks*, *Transit Stations* for public transport hubs, *Workplaces*, and *Residential* which estimates stay-at-home changes. The changes, for a given place on a given day, are reported in comparison to a baseline value, i.e., the median volume for the same day of the week computed across a 5-week

[8] `https://www.mediawiki.org/wiki/ORES/Articletopic`
[9] `https://www.google.com/covid19/mobility`

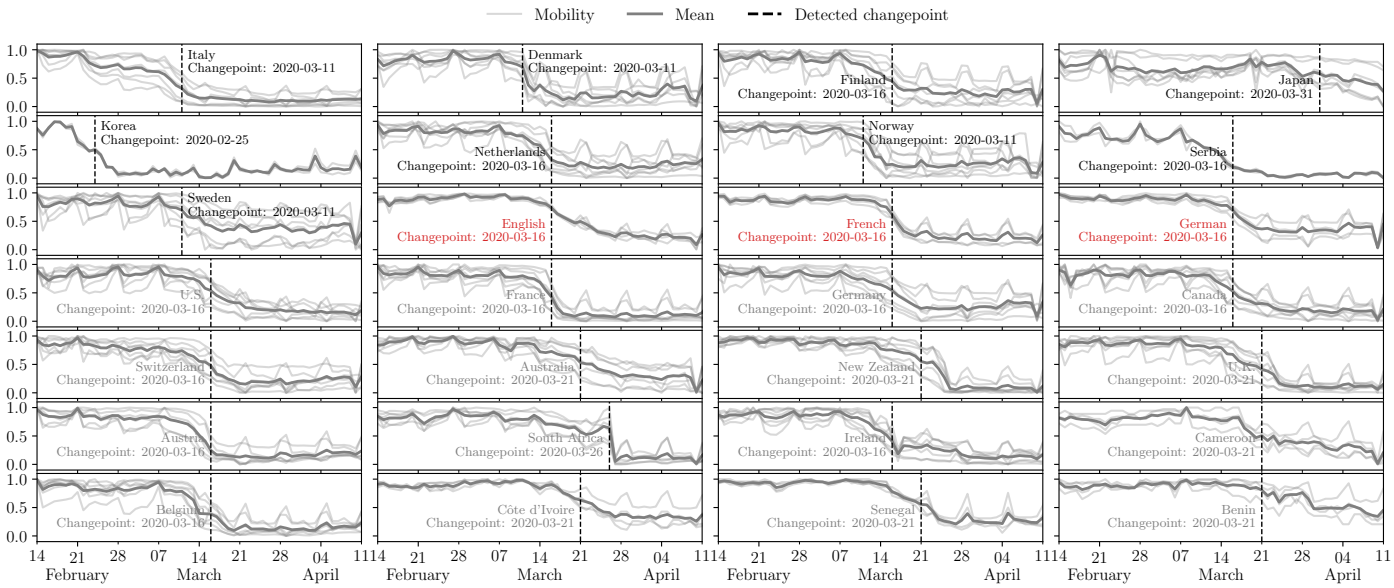


**Figure 8: Mobility data with changepoint detection applied to all countries considered.** In red are the changepoints resulting from an aggregation of several countries, in grey, countries used for the aggregation, and, in black, countries corresponding to a single language.

period between January 3rd and February 6th. Similarly, Apple reports[10] relative changes compared to a baseline volume measured on January 13th along 3 dimensions of mobility: *Driving*, *Walking*, and *Transit*. Combined, these represent 9 types of time-series capturing mobility behavior.

**Dealing with language / country mismatch.** Some languages like English, French or German cannot easily be matched to one particular geographical area. For them, we collect the largest countries in terms of native speakers for which mobility data was available. We then produce an aggregate of the mobility data across countries weighted by the percentage of native speakers of the language in each country. For English, we aggregated: United States (68.9%), United Kingdom (16.1%), Canada (5.8%), Australia (5.4%), South Africa (1.5%), Ireland (1.2%), and New Zealand (1.1%) For French, we aggregated: France (62.9%), Canada (10.1%), Cameroon (9.2%), Belgium (7.9%), Senegal (4.3%), Benin (3.8%), and Switzerland (1.8%). For German, we aggregated: Germany (87%), Austria (8.7%), and Switzerland (6.3%)

**Changepoint detection.** Changepoint detection, is the task of identifying state changes in time-series. We can benefit from the large literature on offline changepoint detection to identify the effective time of lockdown based on changes in mobility time-series.

These techniques usually rely on two components: (i) a cost-function assessing the quality of a particular signal segmentation and (ii) a technique to search the space of possible segmentations, guided by the cost-function [16]. For more details, we refer to the review of [17]. In this work, we considered the cost-functions and search algorithms available as part of the `ruptures` package. [11]

Several different design choices can be made: (i) different subsets of mobility time-series can be selected and (ii) different changepoint algorithms can be employed We ran the pipeline with many different parameters and report the results in Table 2. The standard deviations are due to changing parameters and are small, often below 1 day. Indeed, for most countries, the changes in mobility are very clear, and different methods largely agree.

Finally, Fig. 8 summarizes the mobility data and detected changepoints.

## A.3 Multiplicative effect of mobility decrease across topics

In Fig. 7, we present topic-specific effect of mobility decrease based on a model that is a variant of cf. Eq. 3, where we apply logarithm of the dependent variable $y$ ($R^2 = 0.88$). This allow as to compare topics of different sizes.

In that way, for each topic, we estimate the multiplicative effect of mobility decrease, that is the strongest for VIDEO GAMES and BOOKS. Those are topics that get a small fraction of pageview share, compared to, for example BIOLOGY, that has the highest effect in the additive model Fig. 5.

[10] `https://www.apple.com/covid19/mobility`
[11] `https://github.com/deepcharles/ruptures`