A scientometric overview of CORD-19

Giovanni Colavizza*.[1], Rodrigo Costas[2], Vincent A. Traag[2], Nees Jan van Eck[2], Thed van Leeuwen[2], and Ludo Waltman[2]

[1] *University of Amsterdam*
[2] *Centre for Science and Technology Studies (CWTS), Leiden University*

April 17, 2020

Abstract

As the COVID-19 pandemic unfolds, researchers from all disciplines are coming together and contributing their expertise. CORD-19, a dataset of COVID-19 and coronavirus publications, has recently been published alongside calls to help mine the information it contains, and to create tools to search it more effectively. Here, we focus on the delineation of the publications included in CORD-19, and analyse this delineation from a scientometric perspective. We find that CORD-19 contains research not only on COVID-19 and coronaviruses, but on viruses in general. Publications from CORD-19 mostly focus on a few, well-defined areas, including: coronaviruses (primarily SARS, MERS, COVID-19); public health and viral epidemics; the molecular biology of viruses; influenza and other families of viruses; immunology and antivirals; methodology (testing, diagnosing, clinical trials). CORD-19 publications published in 2020, especially focused on topics of pressing relevance (spread, infection, efficacy of counter-measures), are disproportionately popular on social media. While we fully endorse the initiative that led to CORD-19, we also advise to consider its relatively broad content critically.

COVID-19, CORD-19, Coronavirus, Scientometrics, Bibliometrics.

Introduction

The COVID-19 pandemic is attracting the attention of the global scientific community. Medical research on the virus and on the management of the crisis from an epidemiological and healthcare point of view has full priority. Furthermore, many research communities, funding agencies and third-parties are taking action to support the fight against the pandemic with their own expertise and

*Corresponding author, g.colavizza@uva.nl

resources. Multidisciplinary open collaboration will prove instrumental to fight the current and future pandemics [6, 27].

Several initiatives have been taken to share COVID-19-related scientific research as openly as possible, from public, private and non-profit organizations. It has been readily recognized that health crises are also information crises [32, 10, 25, 17]. For example, the World Health Organization maintains a list of relevant research updated daily [33], as well as a portal to provide information to the public [3], similar to the the European Commission [4]. Publishers are opening access to relevant publications, and some open-access publishers have lifted publishing fees for the same reason.

Another initiative is the release of the COVID-19 Open Research Dataset (CORD-19) [2]. CORD-19 is a growing, weekly-updated dataset of COVID-19 publications, capturing new as well as past research on "COVID-19 and the coronavirus family of viruses for use by the global research community."[1] The reason to release this dataset is "to mobilize researchers to apply recent advances in natural language processing to generate new insights in support of the fight against this infectious disease." The initiative has the backing of the US White House [1] and is a partnership of several institutions including the Chan Zuckerberg Initiative, Georgetown University's Center for Security and Emerging Technology, Microsoft Research, the National Library of Medicine of the National Institutes of Health, and Unpaywall. CORD-19 is released together with a set of challenges hosted by Kaggle, mainly focused on automatically extracting structured and actionable information from such a large set of publications. The release of this dataset is a positive call for action directed towards the natural language processing, machine learning and related research communities. This call has been taken up: For example, the ACL conference has announced an emergency NLP COVID-19 workshop which mentions the CORD-19 dataset on its call for papers[2] and a TREC-COVID challenge has been announced on CORD-19[3].

In order to contribute to an informed use of the CORD-19 dataset, we here present a scientometric analysis of its contents. In particular, we raise some questions on the subject matter delineation of CORD-19. Field or subject delineation is a complex task, which usually benefits from an interplay of information retrieval, or search-based approaches, and bibliometric mapping [16, 34, 18]. We compare the coverage of CORD-19 to the Web of Science and find that CORD-19 on the one hand contains more than just research on COVID-19 and coronaviruses, while on the other hand, it might be missing some content. We then provide an overview of the research it contains, enriching the CORD-19 dataset with data from Dimensions [15] and using both citation analysis and text analysis. Our results show that the CORD-19 corpus broadly focuses on biomedical research on viruses and related health issues, and the articles it contains are quite heterogeneous. CORD-19 contains a core of research directly on COVID-19 and coronaviruses, but in addition it contains many articles on

[1] For coronaviruses, see [13].

[2] `https://www.nlpcovid19workshop.org`.

[3] `https://ir.nist.gov/covidSubmit`.

related yet distinct streams of virus research, such as on influenza, molecular biology and public health. Secondly, we discover three periods in the accumulation of literature in this corpus: a pre-SARS (2003) period, a post-SARS period and the current pandemic (2020). We also present a brief analysis of Altmetric [24, 21] data related to the papers in the corpus. We find that recent research (2020) is disproportionately represented in social media and news by Altmetric indicators, especially on Twitter. We conclude by suggesting a critical stance when using CORD-19, and by proposing some directions for future work.

To facilitate the analysis and use of the CORD-19 corpus, we release our code (see SI). In combination with valid access to Dimensions, Altmetric, the Web of Science and Twitter, our results can be replicated. Finally, we underline that we are not domain experts. We invite experts to improve the interpretation of our results, and rectify it wherever necessary.

The CORD-19 publication dataset

The CORD-19 publication dataset contains over $47,000$ articles, of which $36,000$ are equipped with full text (April 4, 2020). The dataset collects publications from the following sources:

- PubMed's PMC open access corpus, via the query:

```
"COVID-19" OR Coronavirus OR "Corona virus"
OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV"
OR "Severe Acute Respiratory Syndrome" OR
"Middle East Respiratory Syndrome".
```

- COVID-19 research articles from the WHO [5].
- bioRxiv and medRxiv pre-prints using the same query as in PMC.

The relatively low number of publications prior to the SARS 2003 outbreak is followed by a steady increase in the number of publications up to 2020, when the growth accelerated even more substantially (Figure 2). The top 20 journals by number of publications are given in Figure 1, highlighting how the corpus is composed of publications from varied sources. The Journal of Virology and PLOS ONE stand out, in particular. The main contributors to the dataset are PMC and Elsevier (Figure 12b), which have made a large part of their relevant literature openly available. The availability of full texts is relatively high, even if not yet complete, and is proportionally stable over time (Figure 12a). This is because, even though most publishers and journals have opened up their publications, some have not yet done so. In particular, the Journal of Virology (top by number of papers in CORD-19), the Journal of Clinical Microbiology and the Journal of Biological Chemistry provide almost no article in full text to this date.

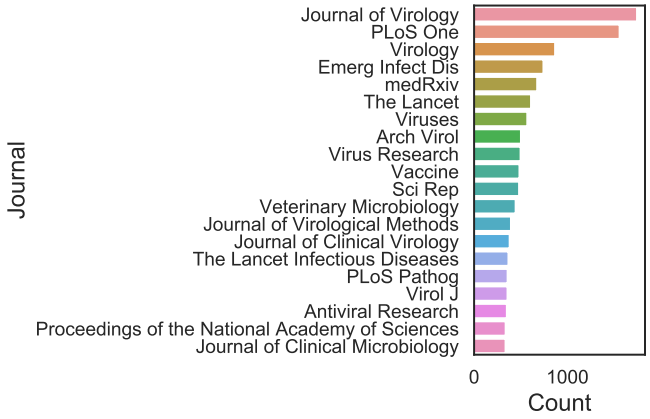
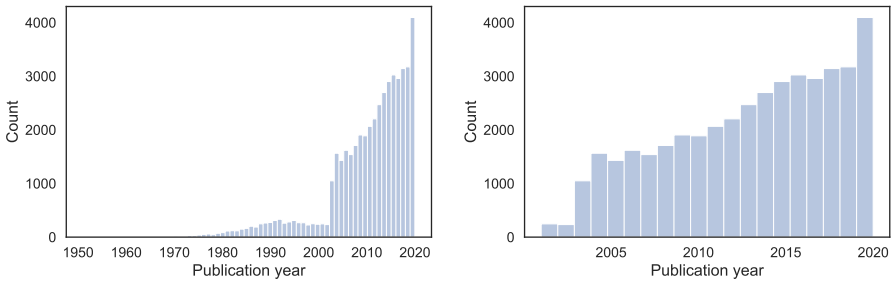


Figure 1: Top 20 journals by number of articles.



(a) Overall.

(b) Since 2000.

Figure 2: Publication years of the CORD-19 papers.

Using Dimensions data, we study the number of citations to papers in the corpus (Figure 13). We can observe how the literature in this corpus gained more attention from the early 2000, and how some publications from 2020 are already accumulating citations at an extreme rate. The most cited articles include: "Epidemiology and causes of preterm birth" by Goldenberg et al. (2008, The Lancet), "Global trends in emerging infectious diseases" by Jones et al. (2008, Nature), "Guidelines for the use and interpretation of assays for monitoring autophagy" by Klionsky et al. (2016, Autophagy), "Biology of Natural Killer Cells" by Trinchieri (1989, Advances in Immunology). The Journal of Virology is the most cited journal in the corpus (in terms of absolute number of received citations from other papers in CORD-19, according to Dimensions data), followed by The Lancet, PNAS and PLOS ONE. Lastly, we consider the Fields of Research (FOR) categories provided by Dimensions, which cover

all areas of research from the Australian and New Zealand Standard Research Classification (ANZSRC).[4] Except for about $10,000$ articles without any classification, we find that, as expected, the Medical and Health Sciences cover most ground, followed by Biological Sciences (Figure 14). The second level FOR classification shows the presence of several sub-areas of these two top level fields, Medical Microbiology being the largest. More details and plots are provided in the accompanying repository.

Possible issues in the delineation of CORD-19

The search terms used to find CORD-19 publications were used to perform a search within the CWTS in-house version of the Web of Science (WoS) [7].[5] We used the following WoS citation indices: Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, and Conference Proceedings Citation Index. Other WoS citation indices were not used, because we do not have access to them. We performed the search in the titles, abstracts, and author keywords of the publications in WoS. Abstracts and keywords are available from 1991 onward. CORD-19 and WoS coverage in publications can then be compared using their DOI or PubMed identifier (PMID). In Figure 3 we show the trends in research output on COVID-19-related diseases. The green line represents the CORD-19 dataset. While this dataset goes back to 1951, for reasons of clarity we only show trends from 1980 to 2019 ($n = 42,496$, excluding an additional $4,095$ publications from 2020 and 403 from before 1980). The red line represents the set of publications collected from WoS using the procedure we just described ($n = 6,958$). From now on, we refer to this search procedure as 'strict'. Finally, the blue line shows all publications from CORD-19 also available in WoS ($n = 33,970$), including those in the red line.

Some differences between the coverage of CORD-19 and WoS are in themselves not surprising: CORD-19 relies on PubMed and pre-print repositories, and has a mono-disciplinary focus including more recent research, while WoS is a multidisciplinary service. However, as we further detail in Figure 4, within the CORD-19 publications found in WoS we can distinguish between the set of $6,958$ papers that carry CORD-19 search terms in their titles, abstracts, and keywords (in red in Figures 3 and 4), and a set of papers which do not ($n = 33,970 - 6,958 = 27,012$) (in purple in Figure 4; 'not strict'). If we assume that authors use the titles, abstracts, and keywords of their publications to indicate the main focus of their research (by explicitly using terms such as "corona virus", "SARS-CoV", "MERS-CoV", or similar), many CORD-19 publications do not clearly focus on COVID-19 and coronaviruses. Presumably, many such publications are more indirectly related to core COVID-19 and coronavirus research.

We observe an increase of publications on COVID-19-related research in the CORD-19 dataset, which is mirrored in WoS (Figure 3). This trend starts

[4] `https://app.dimensions.ai/browse/categories/publication/for`.

[5] Elsewhere in this paper, we use data from Dimensions instead of WoS. Because we do not have access to abstracts in Dimensions, we use WoS data in this section.

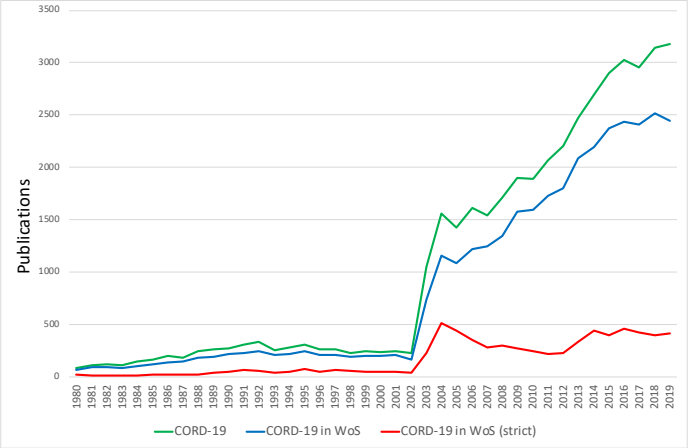


Figure 3: Publication trends in CORD-19 and WoS, 1980-2019.

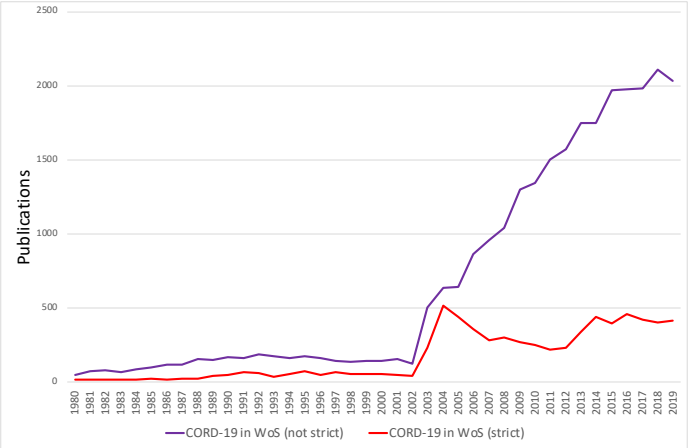


Figure 4: Publication trends in WoS, 1980-2019, distinguishing between results using the 'strict' search procedure.

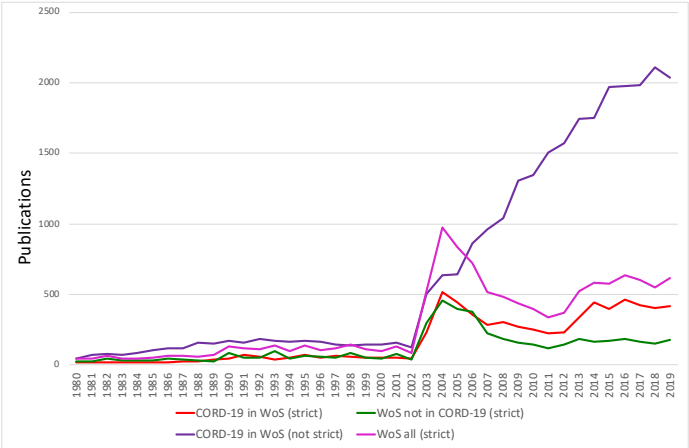


Figure 5: Publication trends in WoS, 1980-2019, highlighting possible COVID-19 and coronavirus research present in WoS and missing from CORD-19.

between 2002 and 2004, with the first SARS outbreak. This first increase in publications is followed by a sustained growth over the years, with some minor fluctuations. However, the set of WoS 'strict' publications, growing between 2002 and 2004, is decreasing afterwards, only to increase again with the MERS outbreak in 2012 (Figure 3). A future direction of research will be to study this apparently widening gap between research directly on COVID-19 and coronavirus and related research.

In Figure 5, we consider publications that are linked to WoS from the CORD-19 dataset. The red and purple lines identify CORD-19 publications from WoS with a 'strict' (red) and 'not strict' (purple) search procedure, respectively. The other two lines represent the set of all publications that are collected from WoS by using CORD-19 search terms in the titles, abstracts, and keywords of publications. The dark green line represents papers that are in WoS, and not in the CORD-19 dataset, but carry the relevant terms in their titles, abstracts, and keywords ($n = 4,830$). Of these publications which are in WoS and should have been in CORD-19, according to search criteria, $3,863$ have at least a DOI or a PMID. All WoS publications matching the CORD-19 search criteria in their titles, abstracts, and keywords are represented by the pink line ($n = 6,958 + 4,830 = 11,788$).

From this analysis we can draw two preliminary conclusions. Firstly, only part of the the CORD-19 dataset concerns publications explicitly related to coronavirus research, while other papers are more indirectly related. Secondly, we identified a set of COVID-19 and coronavirus publications in WoS ($n = 4,830$) that are likely not included in CORD-19. These considerations highlight the need to further qualify the delineation of COVID-19 and coronavirus research that is embedded in CORD-19.

Analysis

We conducted the following analyses on the CORD-19 dataset: a term map and a topic modelling analysis using CORD-19 titles and abstracts; a citation network analysis using Dimensions citation data; an altmetrics analysis using Altmetric data. The purpose of the former two analyses is to cluster CORD-19 publications and further clarify the contents of the dataset. For our analyses, we focus on 46,994 articles out of the 47,351 available in CORD-19, filtering out 53 articles without any known identifier (DOI, PMID or PMCID) and a further 304 due to duplicates.

Term map

To get an accessible high-level overview of the CORD-19 dataset, we used VOSviewer [28] to create a so-called term map of the publications in this dataset. The titles and abstracts were provided as input to VOSviewer, by concatenating them into a single string. Using text mining algorithms of VOSviewer, we included 1,923 most relevant terms of the titles and abstracts. A term is defined as a sequence of nouns and adjectives ending with a noun. Only terms occurring in at least 55 publications were considered. Plural terms were converted to singular. For each pair of terms, VOSviewer counted the number of publications in which the terms both occur in the title or abstract. In this way, a co-occurrence network was obtained, indicating for each pair of terms the number of publications in which they occur together.

This co-occurrence network was visualized in a term map, in Figure 6. The map shows the 1,923 terms included in the network. The size of a term reflects the number of publications in which the term occurs. The proximity of two terms in the map approximately indicates the relatedness of the terms based on their number of co-occurrences. In general, the closer two terms are located to each other, the stronger they are related. This means that groups of terms located closely together in the map usually can be interpreted as topics. The horizontal and vertical axes have no special meaning. In Figure 6, labels are shown only for a selection of the terms. In order to avoid overlapping labels, for many of the less prominent terms no label is shown. The term map can also be explored interactively online (`https://bit.ly/3aijpV3`), including all the overlays used in this article and more. The labels of the less prominent terms can then be made visible by zooming in on specific areas in the map.

Colors are used to present additional information in the term map. In Figure 6, for instance, the color of a term reflects the average publication year of the publications in which the term occurs. In this way, temporal trends can be made visible. The term map shows a clear divide between biologically focused topics in the left area in the map and clinically focused topics and health research in the right area (for similar findings, see [29]). As shown by the colors of the terms, the topics in the right area in the map received a lot of attention in more recent years, while the topics in the left area received attention mostly in earlier years. In the next sections, we use this term map to illustrate some of

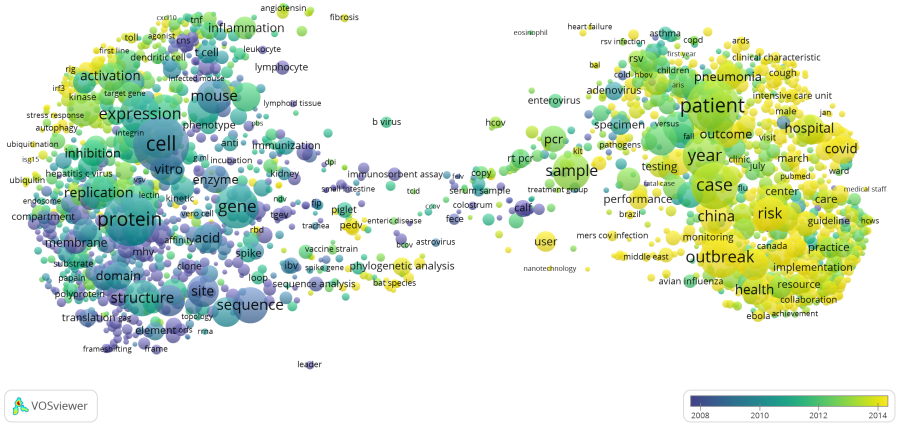


Figure 6: Term map highlighting temporal trends in CORD-19 research. Also compare with topic words in the SI.

the analyses we present.

Topic modelling

We conducted a topic modelling analysis, making use of the titles and abstracts available from the CORD-19 metadata, by concatenating them into a single string similarly to what we did for the term map. Of the $46,994$ articles, $8,349$ have no abstract. We then applied a pre-processing pipeline using ScispaCy's `en_core_sci_md` model [20] to convert each document into a bag-of-words representation, which includes the following steps: entity detection and inclusion in the bag-of-words for entities strictly longer than one token; lemmatisation; removal of (isolated) punctuation and stopwords; inclusion of frequent bigrams.

We started by training a Latent Dirichlet Allocation (LDA) model [9], using `gensim`'s implementation [22] and 15 topics. From a topic coherence analysis [19], we found 15 to 25 to be a good value for the number of topics. We decided to remain on the lower end to facilitate the interpretation of results. We further filtered out tokens which appear fewer than 10 times or in more than half of the documents. Lastly, we verified our results using a Correlated Topic Model [8],[6] which reduces to a Structural Topic Model without covariates [23]. This model explicitly captures topic correlations, and confirms the topic structure we found with LDA, as well as the topics' temporal unfolding. More details are provided in the accompanying repository. We report results of the LDA model in what follows.

The top 20 words per topic are given in the SI. In Figure 15, we show the

[6] Using `tomotopy`, `https://bab2min.github.io/tomotopy`.

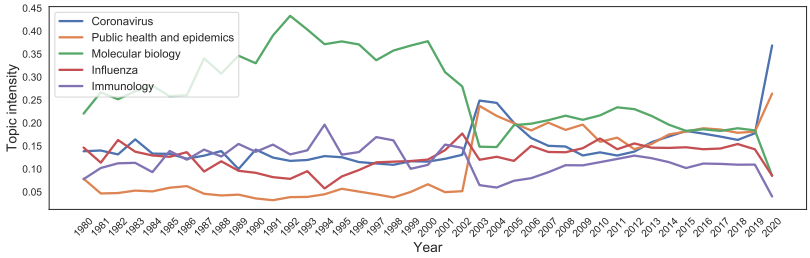
yearly topic intensity, from 1980 to 2020 included. Three periods in the accumulation of literature seem to emerge. Pre-SARS research mainly focused on immunology (topic 1) and molecular biology work on viruses (topics 2, 13), rotaviruses (topic 5), and the genetics of viruses (topic 8). Coronavirus-specific research is mainly covered in three topics, one related to SARS (topic 6), one related to MERS (Topic 9), and one related to the clinical care of coronavirus patients and the more recent COVID-19 (topic 10), see Figure 16. Lastly, another prominent topic started around 2003 and is related to the management of epidemic outbreaks (topic 0). Presently, in 2020, new research is dominated by topics 0 and 10, hence epidemics, outbreak management and COVID-19, especially in view of dedicated patients clinical care. From this first analysis, we observe how coronavirus research seems to be produced in bursts following outbreaks instead of following a more steady progress. Most of the recent and ongoing research at the moment appears to focus on the management of the current crisis.

In order to discuss topics in CORD-19 at a higher level of granularity, we grouped the identified topics into *macrotopics*, all related to research on viruses and their effects, as follows:

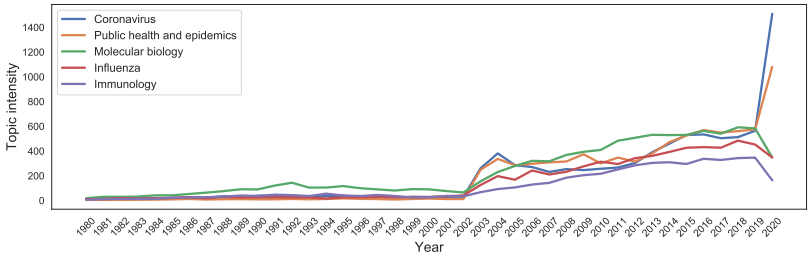
- "Coronavirus": topics 6, 9, 10;
- "Public health and epidemics": topic 0;
- "Molecular biology": topics 2, 8, 13;
- "Influenza": topics 4, 11, 14;
- "Immunology": topic 1;
- "Rotavirus": topic 5;[7]
- "Antivirals": topic 12;
- "Clinical trials": topic 3;
- "Testing and diagnosing": topic 7.

The CORD-19 corpus is dominated by literature on coronaviruses, public health and epidemics, molecular biology and, to a lesser degree, influenza and immunology (five macrotopics above). We plot the relative (yearly mean) and absolute (yearly sum) macrotopic intensities in Figure 7. We see from these plots that the periodization of the CORD-19 dataset discussed above is largely confirmed. The 2003 SARS outbreak generated a shift associated with a rise in publications on coronaviruses and on the management of epidemics. Molecular biology work on viruses remains dominant over time. Lastly, the ongoing COVID-19 pandemic is generating a high proportion of publications primarily on the topics of coronaviruses and the management of epidemics, as is to be expected.

[7] An RNA virus common among infants and children.



(a) Average aggregate; this can be interpreted as the mean topic intensity.



(b) Cumulative aggregate; this can be interpreted as the number of papers per topic.

Figure 7: Macrotopic intensities over time.

Citation clusters

The previous two analyses were based on text, we now turn to a view based on citations to characterized the CORD-19 dataset. We constructed a citation network based on references of all papers included in the CORD-19 dataset, as provided by Dimensions. We not only included references to papers part of CORD-19, but also all other references in this directed citation network. These 'external' references provide additional information regarding the knowledge structure of the CORD-19 dataset. For example, two papers may not be immediately connected via papers in the CORD-19 dataset, but they may have common external references. We only use the giant weakly connected component, which amounts to $903,607$ nodes and $1,985,767$ edges, of which $38,718$ nodes belong to the CORD-19 dataset. This is the citation network that we work with in the remainder of this section.

We cluster the citation network using the Leiden algorithm[8] [26]. Each citation link is weighted as $1/k_i$, where k_i is the number of references of publication i. The inclusion of the 'external' references in the context of clustering is also known as extended direct citation [31]. In this approach, the publications contained in CORD-19 are weighted with a so-called node weight of 1, while the 'external' publications are weighted with a weight of 0 (see [31] for more de-

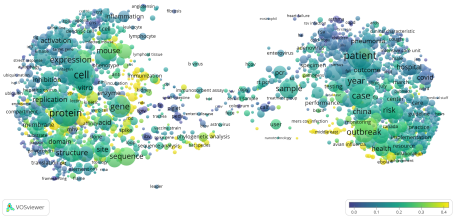
[8]We use the Leiden algorithm implementation from `igraph`.

tails). We cluster the citation network in a hierarchical fashion. The lowest cluster level, which is most detailed, is clustered using a resolution of $2 \cdot 10^{-5}$. We then aggregate the citation network according to this clustering, and then cluster the resulting clustered citation network at a resolution of $1 \cdot 10^{-5}$. We refer to the former clustering solution as the lowest level, and to the latter as the highest level. The 'external' publications help to cluster CORD-19 publications from CORD-19 more accurately. In what follows, we only focus on $38,718$ publications from CORD-19.

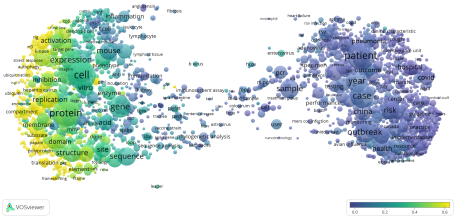
The clustering at the lowest level has 9 clusters that have more than $1,000$ CORD-19 publications each, as shown in Figure 17b. In total, these 9 clusters cover more than 84% of the CORD-19 publications. There are 14 clusters that have fewer than $1,000$ publications but more than 100, covering an additional 11% of the CORD-19 publications. The remaining 5% of publications are scattered across 779 small clusters of just a few publications, of which 651 clusters consist of only a single publication. The largest cluster, consisting of $7,300$ publications, seems to cover publications dealing with various coronaviruses and contains a high proportion of molecular biology analyses, as well as publications dealing with epidemic outbreaks (Figure 8a). The second largest cluster, consisting of $6,176$ publications, has a clear biochemical focus, dealing with various proteins, transcriptions and pathways (Figure 8b). The third largest cluster, consisting of $4,629$ publications, covers clinical studies, revolving around patient care and respiratory conditions (Figure 8c). The fourth largest cluster, consisting of $3,764$ publications, deals with both clinical and epidemiological studies (Figure 8d). This interpretation is largely substantiated by comparing the clustering results to the topic model.

These relatively detailed clusters are hierarchically clustered at a higher level. The largest cluster at the highest level consists mainly of a combination of the first, fourth and fifth largest clusters at the lowest level, and in total it contains $14,325$ publications. This cluster focuses mostly on clinical and epidemiological studies, and outbreak management (Figure 9a). The second largest cluster at the highest level consists mainly of the second, sixth and ninth largest clusters at the lowest level, and in total it contains $10,913$ publications, with a clear molecular biology focus (Figure 9b). These two clusters cover 65% of all publications.

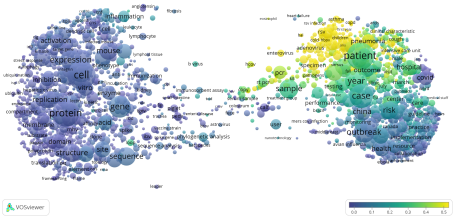
We conclude this section by using topic intensity to qualify the results of the citation network clustering. In particular, clusters can be characterized by the average topic intensity of the publications they contain. We start from the highest clustering level, and focus on the top two clusters by size. We then use macrotopics average intensity to characterize them (Figure 18) and confirm that the first cluster contains publications focused on all coronaviruses, including biomedical and public health research. The second cluster, instead, has a clear molecular biology and immunology focus. We then consider the four largest clusters in the lowest clustering level (Figure 19). While the second cluster is, again, focused on molecular biology and immunology, the first and fourth specialize. The first cluster contains publications on coronaviruses and their molecular biology, while the fourth contains publications on coronavirus



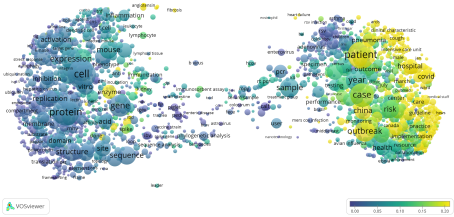
(a) Largest cluster.



(b) Second largest cluster.

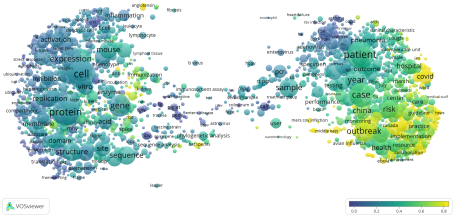


(c) Third largest cluster.

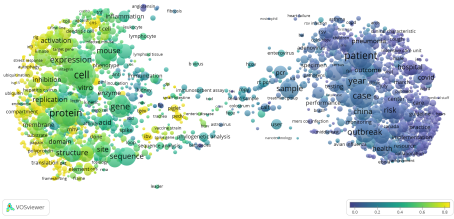


(d) Fourth largest cluster.

Figure 8: Lowest level citation network clustering results. Compare with Figure 19.



(a) Largest cluster



(b) Second largest cluster

Figure 9: Highest level citation network clustering results. Compare with Figure 18.

outbreaks and their public health/epidemiology impact. Lastly, the third cluster focuses on influenza.

This citation based analysis corroborates our earlier findings based on text analyses. Both point to the existence of distinct research areas within CORD-19: coronaviruses, molecular biology research on viruses, public health and epidemics, other viruses (such as influenza) and other related topics (immunology, diagnosing, trials and testing). These areas of research are interrelated, yet also contain specialized information, and accumulated at different rate over time.

Altmetrics

The 46,996 CORD-19 publications have also been explored using Altmetric data, with the aim of describing their reception on social media, paying special attention to the dissemination of the publications across various social media sources. As can be seen in Table 1, a total of 28,146 publications in the CORD-19 dataset (60%) have received some mention in Altmetric. This is a rather high coverage of publications, compared to previous studies [14], that reported an overall coverage of 21.5% of 2012 publications on Twitter, and about 31% for publications in the biomedical and health sciences. This high coverage is even higher when the focus is on the most recent publications (i.e., those published in the early months of 2020), of which over 80% have received some social media mentioning covered by Altmetric.

Table 1: Coverage of CORD-19 publications by Altmetric.

	CORD-19 publications	**In Altmetric**	**Share in Altmetric**
All	46,996	28,146	59.9%
2020	4,096	3,416	83.4%

Table 2 presents a more detailed description of the type of social media events around CORD-19 publications. We selected some of the most relevant sources covered by Altmetric, namely Twitter, blogs, recommendations in F1000Prime, news media mentions, citations in policy documents and citations in Wikipedia entries. Clearly, the most important source is Twitter, which accounts for over 95% of all the (social) media interactions analysed. The second most important source are news mentions (3.6%) and blog citations (0.6%). The observation that news media mentions outperform blogs contrasts with previous studies [14]. This may signal the particular relevance of publications in the CORD-19 dataset for mainstream news media. Another additional characteristic is the recency of the publications being mentioned, particularly on Twitter and in news media, since about 75% of all mentions of publications relate to publications from 2020, while 57% of all mentions in mainstream news media also relate to 2020 publications.

A trend analysis of both publications and tweets (Figure 10) confirms this recency in the uptake of CORD-19 publications. The vast majority of tweets referring to CORD-19 publications concentrate around those published in 2020,

Table 2: Social media events around CORD-19 publications.

	Tweets	Blogs	F1000
All	1,337,036	7,759	700
2020	1,008,839 (75%)	2,989 (39%)	49 (7%)
	News media	**Policy documents**	**Wikipedia citations**
All	50,014	4,596	2,693
2020	28,381 (57%)	163 (4%)	418 (16%)

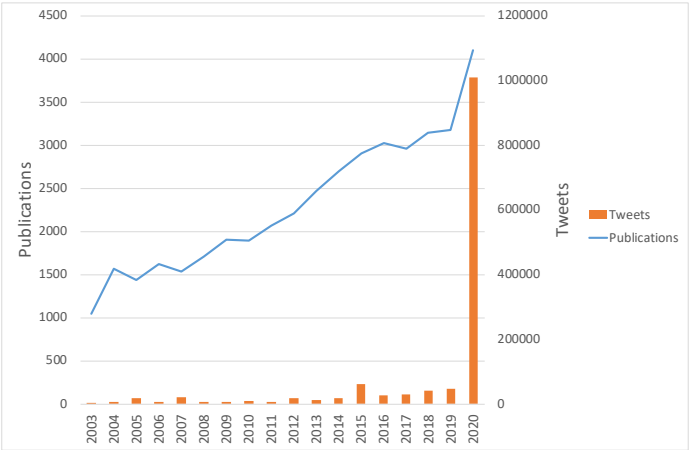
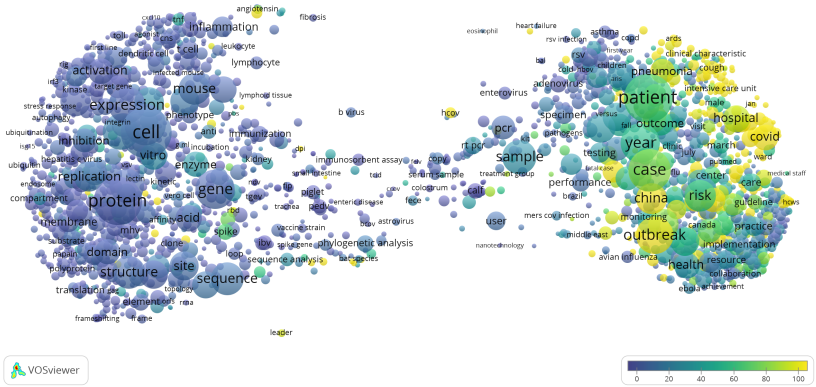


Figure 10: Temporal trend in number of CORD-19 publications and number of tweets received.

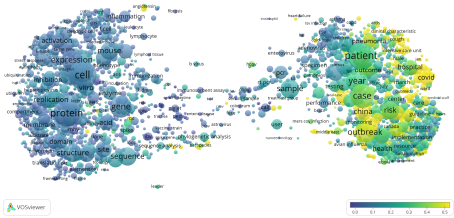
which in practical terms indicates that the social media activity is mostly focused on the ongoing COVID-19 crisis.

Figure 11 presents term maps showing the altmetrics reception of CORD-19 publications. We cover the most immediate social media sources (or 'fast' sources [12]), which provide the earliest signals of the reception of publications. Twitter, blogs and news all present a similar pattern, with a strong orientation towards the most recent COVID-19 publications, well-captured by the fourth largest cluster of the lowest-level citation network clustering (see Figures 8d and 19d).

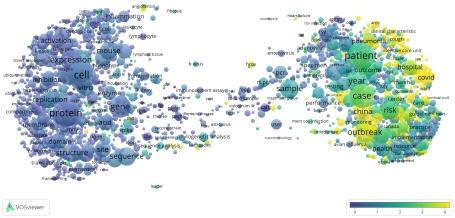
This preliminary altmetrics analysis shows a strong present-day attention for research covered by the CORD-19 dataset. The majority of the tweets, blogs and news mentions are focused on research produced during the current COVID-19 pandemic. This demonstrates the important role of social media, especially Twitter, in discussing COVID-19 research. During a global pandemic like COVID-19, research is subject to high degrees of uncertainty [11, 30]. Rapidly increasing levels of social media activity around topics on which there is little academic consensus may increase the risk of scientific advise being misun-



(a) Twitter.



(b) Blogs.



(c) News media.

Figure 11: Immediate altmetric sources: Twitter, blogs and news media.

derstood or misused by substantial segments of society. Social media analysis can provide tools for identifying and characterizing areas with high levels of social media activity that may be in dissonance with the academic discourse. In future work, we plan to find and characterize these areas of discrepancy in order to inform scientists, science communicators, journalists and the public at large.

Conclusion

We analysed the CORD-19 dataset of publications on COVID-19 and coronavirus research [2]. Comparing the CORD-19 delineation with a WoS-based delineation suggested that on the one hand CORD-19 is broader than just COVID-19 and coronavirus research, while on the other hand, it may also miss some literature that is about COVID-19 and coronaviruses.

We carried out a deeper analyses of CORD-19 in various ways. We created a map of relevant terms extracted from the titles and abstracts of COVID-19 publications. This map confirmed the broad content of the dataset. A topic modelling analysis showed that CORD-19 publications are related more broadly to medical research on viruses, of which COVID-19 and coronaviruses are part. Dominant topics in CORD-19 include research on public health and epidemics; molecular biology; coronaviruses, influenza and other families of viruses; immunology and antivirals; methodology (testing, diagnosing, trials). Furthermore, the topic intensity over time is far from uniform, showing in particular that coronavirus research has followed known outbreaks (SARS, MERS, COVID-19) and that until 2020 this research represented only a small portion of CORD-19.

We performed a citation network clustering analysis using data from Dimensions. Citation network clusters highlight the relative cohesiveness of CORD-19. In line with the textual analyses, the clusters confirm the broad coverage of the dataset. Overall, there seem to be two prominent citation clusters: One that covers research on specific coronaviruses, with a public health and epidemiological focus and another one with a molecular biology focus. Molecular biology research on viruses is, in general, a very prominent component of CORD-19.

Lastly, we considered Altmetric data, in order to gauge how much attention CORD-19 research attracted over time. The current COVID-19 outbreak dominates attention from social media, in particular from Twitter, highlighting the public interest for scientific results during this pandemic.

Far from constituting a critique of CORD-19, our work acknowledges that research on viruses, and coronaviruses specifically, does not exist in a vacuum. Delimiting research on a certain subject matter requires difficult choices that inevitably involve a certain degree of arbitrariness. We praise the breadth and relative coherence of CORD-19. This dataset rightly merits attention and is useful to allow many researchers to engage with the topic. Nevertheless, we also suggest that critical awareness is required when using CORD-19, as our results demonstrate that its contents cover a broad set of topics. Different subsets of CORD-19 should be used for specific purposes, for example for making historical

analyses on funding of specific research topics or for automatically extracting structured information. We exemplified some approaches to segment CORD-19 in various ways. In addition, we also hinted at possible missing content in CORD-19. Some missing content is 'by design', as CORD-19 relies on PubMed, bioRxiv and medRxiv and does not consider, for example, notable pre-print servers such as arXiv, ChemRxiv, JMIR Preprints, SSRN Electronic Journal, Research Square and PsyArXiv. Some missing content is less clearly motivated, as the use of the same search criteria in WoS allowed us to surface published research not currently included in CORD-19. This is an issue that deserves deeper scrutiny.

Clearly, there are many areas for future COVID-19 work by the scientometric community. We conclude by suggesting three areas in particular. Firstly, there seems to be a need for a comprehensive mapping of COVID-19-related research. A multidisciplinary map of COVID-19-related research, considering diverse disciplinary perspectives and information needs, will be useful to surface relevant research, also outside the biomedical domain. Secondly, CORD-19 provides a virtuous example of open data sharing. The scientometric community can contribute by creating and maintaining additional datasets on COVID-19 research. Thirdly, as shown by our results, there is a lot of social media attention for COVID-19 research. Indeed, the role of information, and especially reliable scientific information, has been central to the unfolding of the current pandemic [33]. Consequently, a relevant area for future work is to better understand the mechanics of online scientific information diffusion, using altmetrics and other data sources. This line of work has the potential to provide valuable information to experts and governments during the current and future pandemics.

Data availability

Most of our analysis can be replicated using code and following the instructions given in the accompanying repository: `https://github.com/CWTSLeiden/cwts_covid`. Please note this repository is ongoing work: things might not be perfect. We welcome contributions and suggestions, ideally by opening an issue or doing a pull request. Analyses based on Altmetric, Dimensions, Twitter and Web of Science data require access to these services.

Acknowledgements

We thank Digital Science for providing access to Altmetric and Dimensions data and for constant support throughout our work. We also thank our colleagues at CWTS for enthusiastic and insightful discussions on how the science studies community can contribute to research on COVID-19.

References

[1] Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset, 2020. Released 2020-03-16. Accessed 2020-04-17. URL: https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset.

[2] COVID-19 Open Research Dataset (CORD-19), 2020. Version 2020-04-03. Accessed 2020-04-05. URL: https://pages.semanticscholar.org/coronavirus-research, doi:10.5281/zenodo.3715505.

[3] EPI-WIN: WHO Information Network for Epidemics, 2020. URL: https://www.who.int/teams/risk-communication.

[4] Fighting Disinformation - Official Sources on COVID-19 - Consilium, 2020. URL: https://www.consilium.europa.eu/en/policies/covid-19-coronavirus-outbreak/fighting-disinformation.

[5] WHO COVID-19 Database, 2020. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov.

[6] Juliet Bedford, Jeremy Farrar, Chikwe Ihekweazu, Gagandeep Kang, Marion Koopmans, and John Nkengasong. A new twenty-first century science for effective epidemic response. *Nature*, 575(7781):130–136, November 2019. URL: http://www.nature.com/articles/s41586-019-1717-y, doi:10.1038/s41586-019-1717-y.

[7] Caroline Birkle, David A. Pendlebury, Joshua Schnell, and Jonathan Adams. Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1):363–376, February 2020. URL: https://www.mitpressjournals.org/doi/abs/10.1162/qss_a_00018, doi:10.1162/qss_a_00018.

[8] David M. Blei and John D. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, 2007. URL: http://projecteuclid.org/euclid.aoas/1183143727, doi:10.1214/07-AOAS114.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. URL: http://dl.acm.org/citation.cfm?id=944937.

[10] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media infodemic. *arXiv:2003.05004 [nlin, physics:physics]*, March 2020. arXiv: 2003.05004. URL: http://arxiv.org/abs/2003.05004.

[11] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. The thematic orientation of publications mentioned on social media: Large-scale disciplinary comparison of social media metrics with citations. *Aslib Journal of Information Management*, 67(3):260–288, May 2015. URL: https://www.emerald.com/insight/content/doi/10.1108/AJIM-12-2014-0173/full/html, doi:10.1108/AJIM-12-2014-0173.

[12] Zhichao Fang and Rodrigo Costas. Studying the accumulation velocity of altmetric data tracked by Altmetric.com. *Scientometrics*, March 2020. URL: http://link.springer.com/10.1007/s11192-020-03405-9, doi:10.1007/s11192-020-03405-9.

[13] Anthony R. Fehr and Stanley Perlman. Coronaviruses: An overview of their replication and pathogenesis. In Helena Jane Maier, Erica Bickerton, and Paul Britton, editors, *Coronaviruses*, volume 1282, pages 1–23. Springer New York, New York, NY, 2015. Series Title: Methods in Molecular Biology. URL: http://link.springer.com/10.1007/978-1-4939-2438-7_1, doi:10.1007/978-1-4939-2438-7_1.

[14] Stefanie Haustein, Rodrigo Costas, and Vincent Larivière. Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLOS ONE*, 10(3):e0120495, March 2015. URL: https://dx.plos.org/10.1371/journal.pone.0120495, doi:10.1371/journal.pone.0120495.

[15] Christian Herzog, Daniel Hook, and Stacy Konkiel. Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1):387–395, February 2020. URL: https://www.mitpressjournals.org/doi/abs/10.1162/qss_a_00020, doi:10.1162/qss_a_00020.

[16] Can Huang, Ad Notten, and Nico Rasters. Nanoscience and technology publications and patents: A review of social science studies and search strategies. *The Journal of Technology Transfer*, 36(2):145–172, April 2011. URL: http://link.springer.com/10.1007/s10961-009-9149-8, doi:10.1007/s10961-009-9149-8.

[17] John P.A. Ioannidis. Coronavirus disease 2019: The harms of exaggerated information and non-evidence-based measures. *European Journal of Clinical Investigation*, page e13222, March 2020. URL: http://doi.wiley.com/10.1111/eci.13222, doi:10.1111/eci.13222.

[18] Douglas Henrique Milanez, Ed Noyons, and Leandro Innocentini Lopes de Faria. A delineating procedure to retrieve relevant publication data in research areas: The case of nanocellulose. *Scientometrics*, 107(2):627–643, May 2016. URL: http://link.springer.com/10.1007/s11192-016-1922-5, doi:10.1007/s11192-016-1922-5.

[19] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, UK, 2011. ACM.

[20] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. 2019. `arXiv:arXiv:1902.07669`.

[21] José Luis Ortega. Reliability and accuracy of altmetric providers: A comparison among Altmetric.com, PlumX and Crossref Event Data. *Scientometrics*, 116(3):2123–2138, September 2018. URL: `http://link.springer.com/10.1007/s11192-018-2838-z`, `doi:10.1007/s11192-018-2838-z`.

[22] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[23] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems (NIPS). Workshop on Topic Models: Computation, Application, and Evaluation*, 2013.

[24] Nicolás Robinson-García, Daniel Torres-Salinas, Zohreh Zahedi, and Rodrigo Costas. New data, new possibilities: Exploring the insides of Altmetric.com. *El Profesional de la Informacion*, 23(4):359–366, May 2014. URL: `https://recyt.fecyt.es/index.php/EPI/article/view/epi.2014.jul.03`, `doi:10.3145/epi.2014.jul.03`.

[25] Briony Swire-Thompson and David Lazer. Public health and online misinformation: Challenges and recommendations. *Annual Review of Public Health*, 41(1):433–451, April 2020. URL: `https://www.annualreviews.org/doi/10.1146/annurev-publhealth-040119-094127`, `doi:10.1146/annurev-publhealth-040119-094127`.

[26] Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, December 2019. URL: `http://www.nature.com/articles/s41598-019-41695-z`, `doi:10.1038/s41598-019-41695-z`.

[27] Jay Joseph Van Bavel, Katherine Baicker, Paulo Boggio, Valerio Capraro, Aleksandra Cichocka, Molly Crockett, Mina Cikara, Alia Crum, Karen Douglas, James Druckman, John Drury, Oeindrila Dube, Naomi Ellemers, Eli J Finkel, James Fowler, MIchele Gelfand, Shihui Han, S. Alex Haslam, Jolanda Jetten, Shinobu Kitayama, dean mobbs, Lucy E. Napper, Dominic Packer, Gordon Pennycook, Ellen Peters, Richard Petty, David Gertler Rand, Steve Reicher, Simone Schnall, Azim Shariff, Sandra Sandra Smith,

Linda Skitka, Cass R. Sunstein, Nassim Tabri, Joshua A Tucker, Sander van der Linden, Paul van Lange, Kim Weeden, Michael Jeremy Adam Wohl, jamil zaki, Sean Zion, and Robb Willer. Using social and behavioural science to support COVID-19 pandemic response. preprint, PsyArXiv, March 2020. URL: https://osf.io/y38m9, doi:10.31234/osf.io/y38m9.

[28] Nees Jan van Eck and Ludo Waltman. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, August 2010. URL: http://link.springer.com/10.1007/s11192-009-0146-3, doi:10.1007/s11192-009-0146-3.

[29] Nees Jan van Eck, Ludo Waltman, Anthony F. J. van Raan, Robert J. M. Klautz, and Wilco C. Peul. Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLOS ONE*, 8(4):e62395, April 2013. URL: http://dx.plos.org/10.1371/journal.pone.0062395, doi:10.1371/journal.pone.0062395.

[30] Matthew L. Wallace and Ismael Ràfols. Institutional shaping of research priorities: A case study on avian influenza. *Research Policy*, 47(10):1975–1989, December 2018. URL: https://linkinghub.elsevier.com/retrieve/pii/S0048733318301707, doi:10.1016/j.respol.2018.07.005.

[31] Ludo Waltman, Kevin W. Boyack, Giovanni Colavizza, and Nees Jan van Eck. A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies*, 2020. URL: https://www.mitpressjournals.org/doi/abs/10.1162/qss{_}a{_}00035, arXiv:1901.06815, doi:10.1162/qss_a_00035.

[32] Bo Xie, Daqing He, Tim Mercer, Youfa Wang, Dan Wu, Kenneth R. Fleischmann, Yan Zhang, Linda H. Yoder, Keri K. Stephens, Michael Mackert, and Min K. Lee. Global health crises are also information crises: A call to action. *Journal of the Association for Information Science and Technology*, March 2020. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24357, doi:10.1002/asi.24357.

[33] John Zarocostas. How to fight an infodemic. *Lancet*, 395(10225), February 2020. URL: https://linkinghub.elsevier.com/retrieve/pii/S014067362030461X, doi:10.1016/S0140-6736(20)30461-X.

[34] Michel Zitt. Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation. *Scientometrics*, 102(3):2223–2245, March 2015. URL: http://link.springer.com/10.1007/s11192-014-1482-5, doi:10.1007/s11192-014-1482-5.

SI

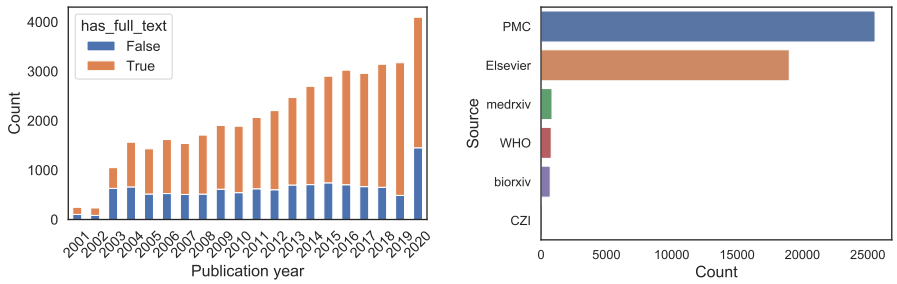
Topic top words

Top 20 words per topic, for the LDA model. We filter out words shorter than 3 characters. Compare with Figure 15 for the topic intensity over time.

- **Topic #0**: "health", "disease", "public", "datum", "model", "outbreak", "public_health", "infectious", "epidemic", "system", "research", "control", "infectious_disease", "study", "country", "review", "Health", "number", "provide", "method".
- **Topic #1**: "cell", "mouse", "infection", "response", "immune", "expression", "disease", "induce", "role", "viral", "gene", "increase", "type", "level", "study", "result", "cytokine", "activation", "lung".
- **Topic #2**: "virus", "cell", "protein", "viral", "replication", "infection", "hepatitis", "mhv", "mutant", "result", "culture", "mouse", "particle", "membrane", "mouse_hepatitis", "virion", "infected", "golgi", "suggest", "show".
- **Topic #3**: "group", "study", "day", "high", "level", "cat", "effect", "significantly", "result", "increase", "rat", "compare", "control", "significant", "low".
- **Topic #4**: "infection", "dog", "hospital", "canine", "transmission", "h1n1", "worker", "contact", "pandemic", "air", "care", "exposure", "2009", "influenza", "healthcare", "control", "mask", "risk", "staff", "room".
- **Topic #5**: "calf", "strain", "isolate", "rotavirus", "bovine", "diarrhea", "sample", "detect", "coli", "study", "analysis", "fecal", "group", "herd", "intestinal", "cattle", "farm", "genotype", "China".
- **Topic #6**: "antibody", "protein", "coronavirus", "sars-cov", "spike", "pedv", "neutralize", "monoclonal", "porcine", "respiratory", "severe", "acute", "epitope", "serum", "mab", "monoclonal_antibody", "severe_acute".
- **Topic #7**: "assay", "detection", "sample", "test", "method", "detect", "virus", "pcr", "diagnostic", "sensitivity", "clinical", "diagnosis", "result", "rapid", "positive", "real-time", "rt-pcr", "specificity", "reaction".
- **Topic #8**: "rna", "sequence", "gene", "genome", "virus", "region", "viral", "analysis", "mrna", "mutation", "protein", "site", "nucleotide", "genomic", "structure", "synthesis", "translation", "transcription", "contain", "genetic".
- **Topic #9**: "virus", "human", "infection", "disease", "animal", "bat", "mers-cov", "respiratory", "host", "specie", "cause", "Middle", "transmission", "pathogen", "coronavirus", "syndrome", "emerge", "viral", "mers", "respiratory_syndrome".

- **Topic #10**: “patient”, “case”, “clinical”, “severe”, “acute”, “disease”, “pneumonia”, “covid-19”, “treatment”, “risk”, “coronavirus”, “respiratory”, “infection”, “day”, “mortality”, “report”, “symptom”, “associate”.
- **Topic #11**: “respiratory”, “infection”, “virus”, “child”, “viral”, “influenza”, “patient”, “study”, “tract”, “pathogen”, “human”, “rsv”, “acute”, “year”, “detect”, “bacterial”, “cause”, “illness”, “clinical”, “common”.
- **Topic #12**: “activity”, “antiviral”, “drug”, “inhibitor”, “treatment”, “protease”, “compound”, “effect”, “inhibit”, “target”, “vitro”, “agent”, “study”, “potential”, “enzyme”, “in vitro”, “inhibition”, “show”, “acid”, “active”.
- **Topic #13**: “protein”, “domain”, “interaction”, “receptor”, “binding”, “membrane”, “peptide”, “target”, “complex”, “structure”, “bind”, “fusion”, “host”, “function”, “human”, “cellular”, “vaccine”, “viral”, “role”, “surface”.
- **Topic #14**: “virus”, “influenza”, “vaccine”, “strain”, “influenza_virus”, “ibv”, “avian”, “vaccination”, “pig”, “infectious”, “challenge”, “chicken”, “prrsv”, “bronchitis”, “swine”, “porcine”, “h5n1”, “result”, “protection”, “supplementary”.

Extra figures



(a) Availability of full text over time. (b) Data sources.

Figure 12: Availability of full text and data sources.

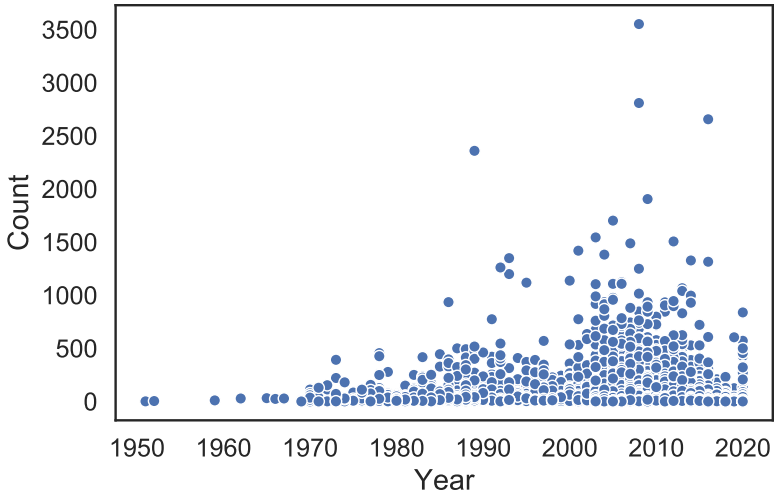
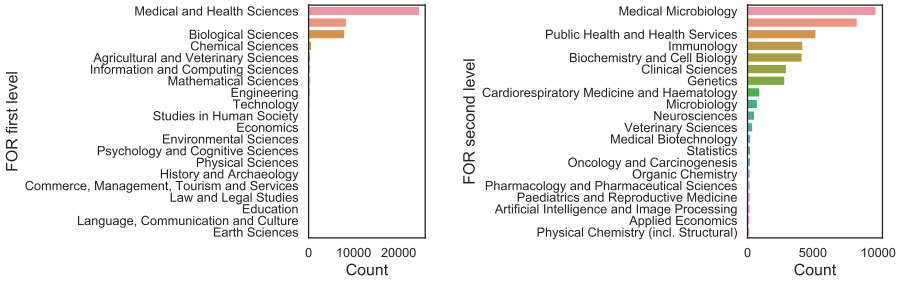


Figure 13: Scatter plot of the number of received citations over time.



(a) First-level categories. (b) Second-level categories.

Figure 14: Categories (FOR Dimensions classification). The empty label account for articles missing a FOR category.

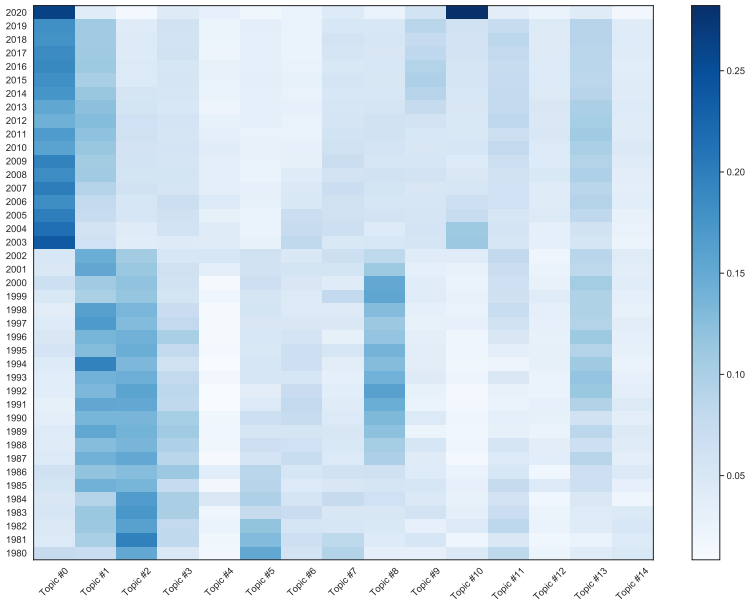


Figure 15: Topic intensity over time, using the LDA model.

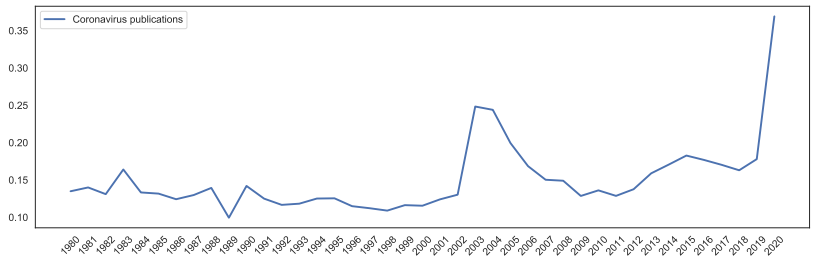
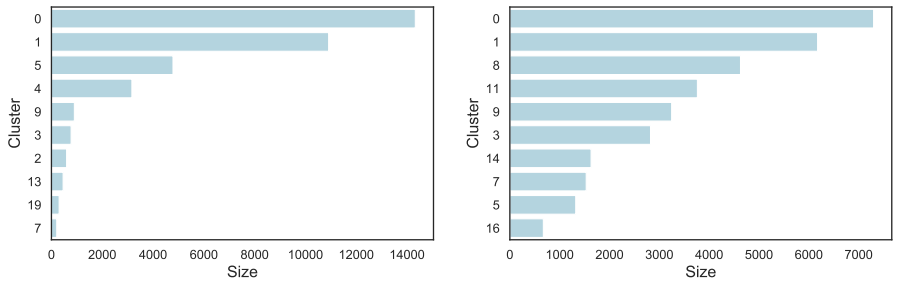


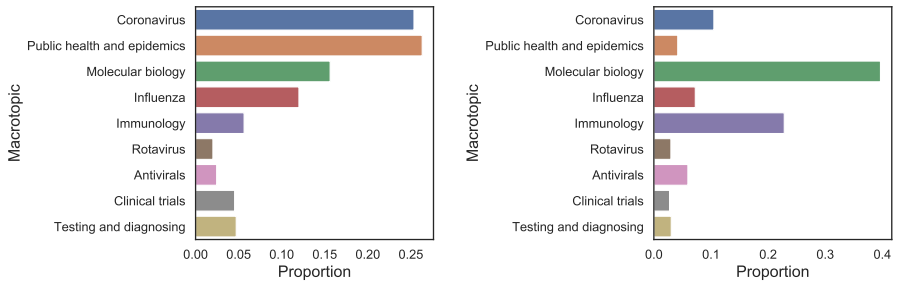
Figure 16: Coronavirus-related topic intensity over time, using the LDA model. This plot considers the document average intensity of topics related to SARS, MERS and coronavirus/COVID-19. Compare with Figure 3.



(a) Highest level clustering: fewer, larger clusters.

(b) Lowest level clustering: more, smaller clusters.

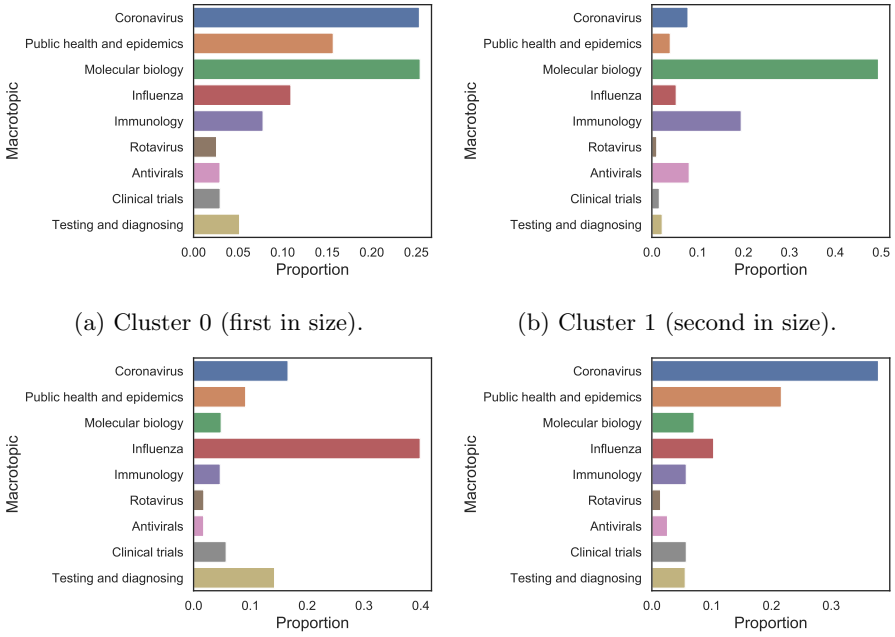
Figure 17: Citation network lowest and highest cluster sizes.



(a) Cluster 0 (first in size).

(b) Cluster 1 (second in size).

Figure 18: Highest citation network clustering (larger clusters).



(a) Cluster 0 (first in size).

(b) Cluster 1 (second in size).

(c) Cluster 8 (third in size).

(d) Cluster 11 (fourth in size).

Figure 19: Lowest citation network clustering (smaller clusters).