Please cite as: Zhang, Y., Cai, X., Fry, C., Wu, M., Wagner, C. 2021, Topic evolution, disruption and resilience in early CoVID-19 research, *Scientometrics*, DOI: 10.1007/s11192-021-03946-7

Topic Evolution, Disruption and Resilience in Early COVID-19 Research

Yi Zhang[1], Xiaojing Cai[2,3], Caroline V. Fry[4], Mengjia Wu[1], Caroline S. Wagner[3, *]

[1]Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia yi.zhang@uts.edu.au, mengjia.wu@uts.edu.au
[2]School of Public Affairs, Zhejiang University, Hangzhou, Zhejiang 310058, China xjcai19@gmail.com
[3]John Glenn College of Public Affairs, The Ohio State University, Columbus, OH 43210, USA
[4] University of Hawai'i at Manoa Shidler College of Business, Honolulu, USA cvfry@hawaii.edu

Corresponding Email: wagner.911@osu.edu

ORCID: 0000-0002-7731-0301 (Yi Zhang); 0000-0001-7346-6029 (Xiaojing Cai); 0000-0002-6874-7608 (Caroline V. Fry; 0000-0003-3956-7808 (Mengjia Wu); 0000-0002-1724-8489 (Caroline S. Wagner).

Abstract

The COVID-19 pandemic presented a challenge to the global research community as scientists rushed to find solutions to the devastating crisis. Drawing expectations from resilience theory, this paper explores how the trajectory of and research community around the coronavirus research was affected by the COVID-19 pandemic. Characterizing epistemic clusters and pathways of knowledge through extracting terms featured in articles in early COVID-19 research, combined with evolutionary pathways and statistical analysis, the results reveal that the pandemic disrupted existing lines of coronavirus research to a large degree. While some communities of coronavirus research are similar pre- and during COVID-19, topics themselves change significantly and there is less cohesion amongst early COVID-19 research compared to that before the pandemic. We find that some lines of research revert to basic research pursued almost a decade earlier, whilst others pursue brand new trajectories. The epidemiology topic is the most resilient among the many subjects related to COVID-19 research. Chinese researchers in particular appear to be driving more novel research approaches in the early months of the pandemic. The findings raise questions about whether shifts are advantageous for global scientific progress, and whether the research community will return to the original equilibrium or reorganize into a different knowledge configuration.

Keywords

COVID-19; topic analysis; science; research and development; international collaboration.

1. Introduction

In all the attention given to research on COVID-19, ample studies have focused on who is working with whom (Banda et al., 2020; Colavizza et al., 2020; Fry et al., 2020; Kokudo & Sugiyama, 2020; Kyhlstedt & Andersson, 2020; Mohamed et al., 2020). In contrast, the literature places much less focus on the type and direction of research during the pandemic. In earlier work, we showed that in the earliest days of the pandemic there was an explosion of research on coronavirus-related topics, and that China and the US led the effort, on their own as well as cooperating actively on COVID-19 research (Fry et al., 2020). Complementing this early analysis with data from additional months, we find that a total of 18,000 papers had been published on coronavirus-related topics between January and the end of June 2020[1]. We find that the number of authors on coronavirus research articles immediately dropped at the onset of the pandemic, and it has continued to drop over the COVID-19 period. This rush to produce relevant research, combined with the observation that the structure of scientific teams has changed during the pandemic, raises the question of how the crisis influences the trajectory of research being conducted.

In this paper, we explore how the crisis affected the trajectory of coronavirus research by viewing the scientific process as a complex system that can be modeled and studied (Contractor et al., 2006; Lee & Monge, 2011; Monge et al., 2008). This approach includes modeling parts of a multilevel construct, as complex systems are characterized by hierarchies (Simon, 1991) or multiples levels (Monge & Contractor, 2003). We consider clusters of topics, and topics themselves, as levels of the hierarchy of the system of coronavirus research. The clusters represent the epistemic organization of fields. Expectations on how topics and clusters of topics are affected by the crisis are drawn from resilience theory, which was developed by ecologists to explain how systems achieve and maintain equilibrium, and how they recover after a catastrophic disruption (Folke, 2006; Holling, 1973; Walker & Salt, 2012). Prior research in this area has identified that a system will attempt to absorb a disturbance and re-organize whilst maintaining a similar structure and function, but at the same time disturbance allows for the emergence of new trajectories and some new features to emerge. By comparing the trajectory of coronavirus research and the stability of scientific topics before and during the COVID-19 crisis, we are able to contribute to a better understanding of the extent to which the research community draws on prior knowledge and re-stabilizes during the crisis, and whether coronavirus research becomes more novel during COVID-19. We expect that, following a disruption, the coronavirus topics will return to core clusters of topics (or 'pillars', strong 'species') and will begin to reorganize around these pillars. We further expect to see some 'fragile' and perhaps less relevant topics and clusters of topics fall off, while other topics reorganize or draw from previous periods in response to the crisis.

In order to test these propositions, we analyze the various levels of the system of coronavirus research before COVID-19 and in the early months of the pandemic. We compare topic clustering of the two time periods - the pre-COVID-19 period (2009 – 2019) and the COVID-19 period (January – April 2020). This paper uses concepts from network analysis, combined with topic evolution and statistical analysis to reveal how coronavirus topics evolve and recombine across the corpus of knowledge to address a critical scientific problem. Specifically, we use data on key terms featured in scientific articles before and during the COVID-19 pandemic to view the evolutionary pathways of topics in order to characterize research trajectories and to see how much prior knowledge fed into early pandemic research. We also explore the impact of the disruption on the knowledge space surrounding coronavirus research. We study the relationships between communities of topics, and the prevalence of different actors across the map of research communities using network analysis and statistical methods.

This paper is organized as follows: Section 2 discusses related work in topic analysis by reviewing previous studies. Section 3 details the data and methodology for this project, outlining the research framework and details about data acquisition and analysis. Section 4 presents the results and empirical insights identified during the study. Section 5 offers discussion and conclusions.

2. Literature Review

A large literature has explored the determinants of the direction of research across a number of different disciplines. Studies have found that the direction of research across communities of scientists can be explained by incentives (Acemoglu & Linn, 2004; Azoulay et al., 2019; Finkelstein, 2004); peers and team composition (Catalini et al., 2020;

[1] Gathered from sources: Scopus, Web of Science, PubMed Central, and Preprints.

Ganguli, 2015) and the availability of supporting infrastructure and tools (Furman & Teodoridis, 2020). Despite this progress, however, this literature is limited in the extent to which it can help predict what would happen in a crisis. During a crisis all of these drivers of research direction change, in addition to a disruption to the underlying system that these researchers are embedded in.

That said, resilience theory could provide some useful lessons on how research trajectories change during a crisis. Resilience analysis compares a system's ability to adjust to disruption and to regain basic functionality after catastrophic events (Gao et al., 2016), and describes how a system persists or changes during a disruption, proposing that *''resilience determines the persistence of relationships within a system and is a measure of the ability of these systems to absorb changes of state variables, driving variables, and parameters, and still persist''* (Holling, 1973, p. 17). The coronavirus research community can be viewed as a network of connections. The nodes self-organize into groups that cluster around epistemic topics of interest, and these topics themselves cluster into relevant groups. Scientific communities and topic clusters have been shown to be complex, self-organizing systems (Borrett et al., 2014; Wagner & Leydesdorff, 2005) and epistemic communities often track with the map of science by reflecting disciplines and subdisciplines (Börner et al., 2012). Events leading to loss of order—in this case, a pandemic—are rarely predictable; in nature, they can cause irreversible damage depending upon resilience and the environment. In a knowledge system, the disruption and resilience of a knowledge community can reveal aspects of knowledge creation that provide insights into dynamics.

Specifically, we expect that the epistemic cluster of coronavirus research will be disrupted and will reassemble to reflect new priorities imposed by the COVID-19 experience. However, not all is lost during a crisis, and a system begins to return to stability over time. In particular some species are more resilient than others during a crisis, and we argue that some topics that are most fit for the changing landscape will be resilient through a crisis whilst others that are less fit will either fall into extinction or reorganize and exhibit novelty. In the same vein, some communities of researchers will be able to stabilize or reorganize better than others, depending on the flexibility in their institutional systems and the underlying knowledge base driving their baseline stable state.

We explore these hypotheses using a combination of topic extraction and network analysis. Topic extraction, clustering, mapping, and analysis is a tool of science, technology, and innovation policy (STIP) analysis (Zhang et al., 2016), pioneered by Allan (2012). Specific software analyzes topics drawn from scientific documents (e.g., research articles, patents, and academic proposals) to trace evolutionary trends in research outputs. Collections of documents, or 'bags of words,' can be tapped to identify trends in technology development, manufacturing processes, materials, and the evolution of research areas (Blei, 2012). Chen et al., (2010) and Ding and Chen (2014) further developed the tools. Lee et al. (2009) and Zhang et al. (2017a) created tracings of the historical pathways of technological innovations.

Topic extraction seeks groupings or patterns, items, and objects from text (Jain, 2010). Clusters of related terms are revealed through clustering algorithms, such as K-means (Jain et al., 1999), latent semantic analysis (Deerwester et al., 1990), and latent Dirichlet allocation (Blei et al., 2003). Granularity can be adjusted based upon the research question: for example, discipline-level topic extraction from a global database can provide the outlines of a discipline. Topic extraction can be combined with other data analytic techniques (e.g., network analysis and natural language processing) or specific bibliometric indicators (e.g., citation/co-citation metrics) to trace topic evolution in the bibliometric literature (Suominen & Toivanen, 2016; Waltman & Van Eck, 2013; Zhang et al., 2018). Novel topics can be difficult to identify if an analyst is limited to the historical list of scientific disciplines (Small et al., 2014), the advantage of this clustering approach is shown in the bottom-up organization of information, obviating the need to bin words into pre-existing categories of science.

Measures of network resilience exist, but these are in early stages of development. Gao et al. (2016) suggested a measure for resilience that improves upon the one-dimensional linear equation common in ecology (Folke 2006). The system is measured in one of the stable fixed points and then again when it loses its resilience and undergoes a sudden transition to a different, often undesirable, fixed point of the equation. Gao et al. (2016) sought to improve upon this static measure by accounting for the dynamic state of a network, its many variables, by offering a multi-dimensional manifold over the complex parameter space characterizing the system. The Gao et al. (2016) method looks very appealing to us, but the technical specifications are difficult to achieve. Thus, we measure network centrality of topics to assess the network structure of coronavirus research in three different states before and during the pandemic. The

three states are the ten years prior to the pandemic, the first three months of COVID-19 research, and the six months prior to this publication, May-October, 2020.

By combining network analysis with topic extraction, in this study we present two sets of networks, one, non-directional networks showing coronavirus research in the ten years before the pandemic and then at two points during the pandemic, where topics are the nodes and connections between topics are the edges. Topics, and clusters of topics, are likened to species in an ecosystem which become disrupted by the pandemic event. In addition, we use directional networks to show topic evolution over time among, where nodes are topics and edges are directional, evolutionary relationships. We examine both networks and complement this with statistical analysis of changes in topics before and after the pandemic, for structural change as a result of COVID-19.

3. Data and Methodology

Data

To study the knowledge system around the coronavirus, our study focused on two datasets, one before and one after start of the 2020 pandemic. To allow comparisons, we used a similar data search strategy and database as that used in our pilot study (Fry et al., 2020). One dataset contains articles about coronavirus in the 10 years leading up to the COVID-19 crisis (between January 1, 2009 and December 31, 2019). The second dataset contains articles, notes, letters, and preprints about coronavirus research during the COVID-19 crisis period (between January 1, 2020 and April 23, 2020).

For the topic clusters, we continue the work begun in our pilot study Fry et al. (2020), where we showed the coronavirus research ecosystem before and in the early days of the COVID-19 crisis through examining research topics featured in published artefacts before and during the crisis. These networks use keyword analysis to reveal epistemic communities.

For all of the analysis, we extracted all articles from the Clarivate Web of Science (WoS), Elsevier Scopus, PubMed Central, and Dimensions (including preprint servers: bioRxiv.org, medRxiv.org, and arXiv.org) that contain the following words in the Title/Abstract/Keywords: "COVID-19" OR "2019-nCoV" OR "coronavirus" OR "Corona virus" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome" in the time period analyzed.

Table 1 shows the summary of data collected. The searches produced 33,598 published articles with author-identifiable information across the two periods. Particularly, in the COVID-19 period, 2,147 preprints were pulled. Duplicate articles were eliminated.

Table 1 Data source and publication data.

Number of publications		
Source	Pre-COVID-19 (January 1, 2009 to December 31, 2019	COVID-19 (January 1, 2020 to April 23, 2020)
Scopus	10,012	1,714
Web of Science	7,838	822
PubMed	28,484	4,334
Preprints (BioRxiv/MedRxiv/arXiv)	N/A	2,147
Combined (duplicates dropped)	30,660	6,337
Combined, with topic data	28,543	3,485
Combined, with topic and affiliation data	27,424	3,128

Note: We included preprints in the COVID-19 period because the time pressures imposed by the pandemic crisis propelled ready and open sharing of even initial results, which may help us understand the early response of researchers to the COVID-19 crisis.

While our main analysis incorporates COVID-19 articles produced between January 1 2020 to April 23 2020, we collect additional data on articles published between May and October 2020, using the same search strategy. We use this supplemental medium to longer-term data to generate co-term map (as shown in Fig. 4). Unless otherwise specified, the COVID-19 period refers to January 1, 2020 to April 23, 2020.

Methodology

Fig. 1 illustrates the framework applied in this study of the coronavirus knowledge ecosystem before and during the COVID-19 pandemic. On the left, the triangle represents three distinct levels of knowledge creation whereby a variety of actors drive the knowledge flow through the levels. The first level of clustering of knowledge is at the topic or component level. We expect that topics can then be clustered at a higher level into communities (research area), and then, eventually, into disciplines (not addressed in this paper). Each level can be studied for characteristics.

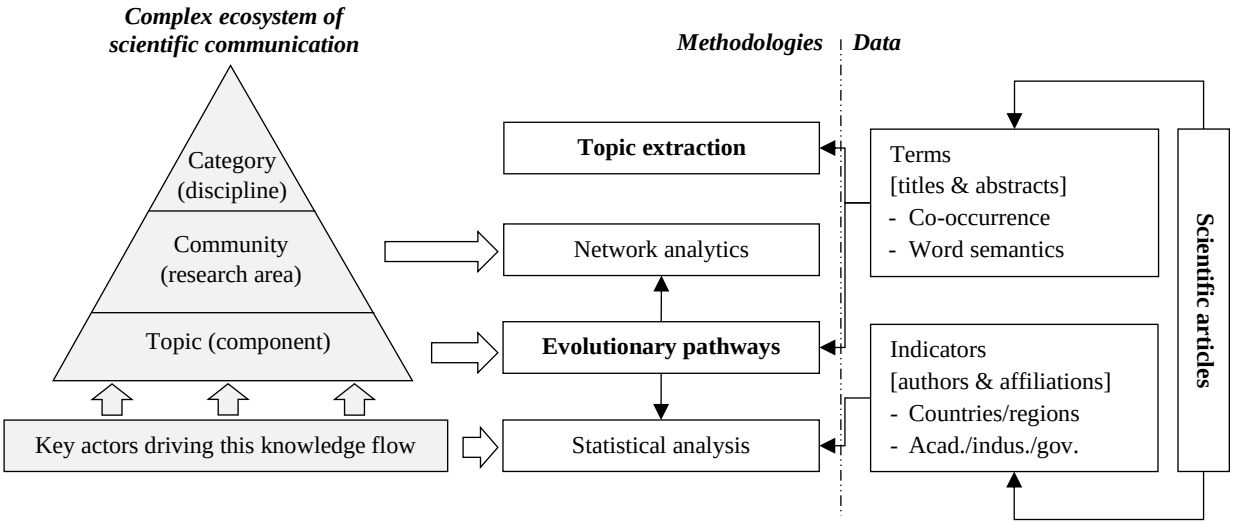

Fig. 1 Research framework

As described in Fig. 1 in the main text, this study focused on terms and bibliographical information collected from scientific articles. The main methodology includes topic extraction and evolutionary pathways. Specifically, topic extraction profiles the technological landscapes of the coronavirus research in the pre- and COVID-19 periods. Evolutionary pathways trace the knowledge flow of the coronavirus research by identifying topics and their relationships over time. Further, we used network analytics to detect research communities from the evolutionary pathways, and then investigated the role of key actors (e.g., affiliations, international collaborations, and research communities) in driving this knowledge flow through statistical analysis. Each of these parts is described below, and in a more detailed Appendix. The aim is to thoroughly understand key research topics in the coronavirus research, to discover how these topics evolve from existing knowledge, and how new knowledge in the COVID-19 period is created.

(1) Data pre-processing

Data pre-processing creates the basic information for the analyses of topic mapping and evolutionary pathways. Using titles and abstracts from the articles in the datasets described above, we conducted two distinct data pre-processing functions on the dataset: First, we applied a natural language processing (NLP) function, integrated in VantagePoint[2] Software, to retrieve terms (i.e., multi-word phrases) from the combined field (titles and abstracts), and then a term-clumping process (Zhang et al., 2014) to identify core terms by removing noise and consolidating synonyms. These become the input in an evolutionary pathways phase. Second, we applied the Word2Vec model (Mikolov et al., 2013) to the raw text of the combined field and generated phrase vectors by matching core terms and word vectors (each

[2] VantagePoint is a software platform for bibliometrics-based text analytics and knowledge management owned by Search Technology Inc. More details can be found at the website: www.vantagepoint.com.

word is represented by a vector, which is the raw output of the Word2Vec model). This set becomes an input in the topic extraction phase.

In the pre-processing stage, the bibliographical information of articles was collected and would be used as indicators for further statistical analysis. At the article level, we categorized articles by author affiliation types and country of the institution affiliated by the author. Specifically, an article's author affiliations were classified into 'academic,' 'industry,' 'government,' or 'other' based on organization types, using full counting. That is, an article involving both 'academic' and 'industry' affiliations are classified as both academic and industrial. This was done using Clarivate's Incites database which allowed us to match extracted affiliations with Incites organizational names and classifications. Further, according to the countries identified in previous work (Fry et al., 2020), a set of dummies were set for each article indicating the presence of international collaboration where at least two distinct countries/regions of author affiliations, such as Chinese authorship, USA authorship, and China-USA collaboration. These data were used later to map topics to sectors.

(2) Topic extraction

Based on the phrase vectors (above), topic extraction was employed to profile the technological landscape of coronavirus research and to identify key research topics in the pre- and COVID-19 time periods. Here, a "topic" is a set of related core terms, representing specific components, such as technologies, research areas, equipment and materials within the corpus. Topics become the basis for co-word maps such as that shown in Figs. 1-3.

At this stage, an additional analysis was conducted to better define and link terms. In earlier work, Zhang et al. (2018) showed that the incorporation of K-means approaches and word embedding techniques is superior in clustering bibliometric data to earlier methods. In this paper, we further refined the method by introducing an "elbow method" (Jain et al., 1999) which seeks the most local-optimal number of topics in an unsupervised way. We further conducted topic extraction by phrase vectors, which provides a richer solution for knowledge representation than would be seen for individual words. Technical details on this unsupervised K-means approach are described as below.

Step 1: Determine the number of topics k and the maximum times of iteration.

Step 2: Randomly initialize k phrase vectors as the starting centroids C of k topics.

Step 3: Assign each phrase vector v to its nearest centroid using cosine similarity maximization (Salton & McGill, 1986), see Equation (1)

$$Cosine\ Similarity(v, C) = \frac{v \cdot C}{\sqrt{v \cdot v} \cdot \sqrt{C \cdot C}} \quad (1)$$

Step 4: Recalculate every centroid by averaging all allocated phrase vectors, see Equation (2)

$$C_i = \frac{1}{Num_i} \sum_{j=1}^{Num_i} v_{i,j} \quad (2)$$

where C_i and $v_{i,j}$ respectively represent the centroid of Topic i and the jth phrase vector in Topic i, and Num_i is the total number of phrase vectors in Topic i;

Step 5: Iterate Steps 3 and 4 until all the k centroids stop moving or the maximum iteration is reached.

The super parameter k of K-means approaches has been criticized for decades because it could sensitively influence the performance of the approaches. Thus, we integrated the 'elbow' method to the above K-means algorithm, which then provides an unsupervised solution for deciding an optimal k in a given interval. The elbow algorithm is described as follows:

Step 1: Provide an interval for setting the number of topics k, and iteratively implement the above K-means algorithm with an incremental k.

Step 2: For each clustering solution, calculate the value of its corresponding distortion $D(k)$ which is expressed by the sum of squared distances from each phrase vector to the centroid of its assigned topic, see Equation (3)

$$D(k) = \sum_{i=1}^{k} \sum_{j=1}^{Num_i} cosine\ distance\left(v_{i,j}, C_i\right) \quad (3)$$

Step 3: Find the local maximum value of $D(k) - D(k-1)$ to numerically identify k that yields the largest decreasing rate in distortion.

The phase of topic extraction produces a list of topics represented by a set of core terms related to the coronavirus research, which provides a clue to understand the technological landscape of related research.

(3) Evolutionary pathways

We applied the scientific evolutionary pathways (SEP) process introduced by (Zhang et al., 2017b) to trace the evolution of scientific topics. The design of the SEP was inspired by an assumption that scientific invention is the recombination of established knowledge (Fleming, 2001; Fleming & Sorenson, 2004), and we then assumed scientific evolution is the result of cumulative changes occurring within established scientific inventions, which could be represented by research topics (e.g., theoretical concepts and technological components). One example is the topic "data mining", which referred to techniques in database management and data warehousing in the 1990s, but it is closely related to machine learning these days, even though database management could be still a part of the topic. That is to say, the evolution of topic "data mining" is reflected by the extension of its feature space (e.g., new features such as "machine learning" were involved) and the change of the distribution of those features (e.g., the proportion of "database management" was decreasing, while that of "machine learning" was increasing). Given the challenge, the SEP algorithm was developed to track scientific evolution by monitoring a topic's feature space and the distribution of these features. Specifically, the connections between evolved topics (e.g., machine learning-based data mining) and their original topics (e.g., database management-based data mining) were defined as predecessor-descendant relationships.

We applied the SEP approach to track the convergence and divergence of research topics on coronavirus research and reveal connections between COVID-19 research and prior knowledge (for technical details on the SEP approach, please see Appendix B). We traced the evolutionary pathways of coronavirus research in the past decades (2009-2020) through topics and their predecessor-descendant relationships, which help us to discover potential knowledge flows and knowledge recombination between COVID-19 and existing research topics.

This analysis is conducted with articles as the core unit of analysis. Articles co-occurring in the same year are grouped in a 'time slice', then, the entire dataset is analyzed as a bibliometric stream. The stream connects topics across time by classifying them into categories based upon whether they have remained within the text corpus without interruption, called "live" topics, or, they dropped out of use, and therefore called "dead" topics. A third option is those topics that dropped out of use but are revived for coronavirus research, called "resurgent" topics. The latter type of topic recalls the 'sleeping beauties' concept defined by Van Raan (2004), who pointed out that some topics fall away but are revived later when needed for scientific explanation.

This process defines that a 'live' topic could be 'dead' if it does not capture new knowledge (i.e., assigned articles) in two sequential time slices. A 'dead' topic may be revived and become 'resurgent' if a new topic shares high similarity with it. We specifically focused on three types of topics:

- "Always alive" topics – topics that were born early (e.g., several years before 2020 – in this paper we specifically chose topics born in 2017 or before) and are always alive and never become 'dead', which may indicate key research areas of the field.
- "Resurgent" topics – topics that are 'dead' but were resurged later, and are alive until the last time slice (i.e., 2020), which may indicate certain resurging interests of the community due to the sudden change of related situations (e.g., new materials and equipment, ground-breaking findings, and the upset of existing knowledge).
- "Emerging" topics – topics that were born recently (i.e., 2020), which may indicate new and influential research areas.

Technically, we constructed a universal feature space for the entire dataset, in which each feature represents one term. Thus, one article could be represented as a vector, in which '1' means the article contains the term represented by that feature, vice versa. Geometrically, we described a topic as a circle, using a 'centroid' (i.e., the mean of all involved articles) and a 'boundary' (i.e., the largest Euclidean distance between the centroid and its involved articles), and the analysis of evolutionary pathways seeks similarity between a current article and centroids of all "live" topics via Salton's cosine (Salton & McGill, 1986). The similarity in the SEP algorithm is the key to monitor the change of topics (in either their feature space or the distribution of their features). We assigned each article to its most similar topic. If an article's Euclidean distance to the centroid of a topic is smaller than its boundary, it will be directly assigned to the topic, indicating the content of this article is closely related to the topic. Or else, this article will be labeled as 'drift', since its content is not exactly the same as this topic, indicating potential evolution might occur in the topic. Then, we moved to analyze the next article.

At the end of each 'time slice', we checked the status of each topic – i.e., set topics as 'live,' 'dead,' and 'resurgent.' For each 'live' topic, we applied the unsupervised K-means approach introduced above to those assigned 'drift' articles and grouped them into certain sub-topics.

We measured the cosine similarity between each sub-topic and two sets of topics - its assigned 'live' topic and all 'dead' topics. If the descendant topic is similar with its assigned one, their relationship is defined as 'predecessor-descendent', or else, the most similar 'dead' topic will be revived and set as 'live', which then becomes a predecessor of the next sub-topic. This is the practice of 'sleeping beauty' detection, in which we semantically evaluated the connections between new knowledge (i.e., sub-topics) and resurgent 'sleeping beauties'.

Then, we labelled a descendant-topic via the term with the highest similarity to all other terms in this topic. If the term has been used before, we would choose the term with the second highest similarity, et cetera. This labelling strategy will select high-frequency terms in early time slices but along with time relatively low-frequency terms will be highlighted. This strategy provides a solution of using a set of labels to comprehensively describe a community, described in Fig. 1 – imaging some high-frequency terms representing basic knowledge in the root and some relatively low-frequency terms at the end representing their follow-up evolution. Due to the use of low-frequency terms, this labelling strategy may result in certain unexpected topics, whose labels could not exactly reflect the main content of their involved articles, because a perfect label for this content has been used by other topics but most of those topics might be their predecessors in the same community. Thus, as given in Fig. 1, the following statistical analysis for measuring the role of key actors emphasizes the community and category level, rather than individual topics.

At the end of each time slice, we updated all 'live' topics by updating their centroid and boundary, and then moved to the next time slice and began the process again.

Results of the SEP approach include a list of topics and their predecessor-descendant relationships as well as the statistical information of each topic, such as labels, descriptive terms, numbers of terms and records, and indicators of 'sleeping beauty' detection (e.g., time of introduction, latency, and resurgence).

The topics were then visualized in a directed network via Gephi (Bastian et al., 2009). In the network, each topic is represented by a node. A directed edge represents the 'predecessor-descendant' relationship between the connected nodes; the weight of an edge reveals the strength of the relationship measured by the cosine similarity. The color of nodes reflects their communities, which are identified using the community detection algorithm integrated in Gephi as "modularity" (Newman, 2006). Since nodes in the evolutionary pathways may represent detailed topics and concepts of COVID-19 research, a community could be considered a group of similar nodes, aligning within the same research areas but with different foci. The size of nodes represents diverse indicators – for example, 1) the importance of a topic, which is defined by the value of term frequency inverse document frequency (tf-idf) analysis, and 2) the role of China, which is calculated by the ratio of articles with at least one Chinese researcher in each topic.

(4) Statistical Analysis

The statistical analysis allows us to investigate the role of key actors driving the knowledge flow of coronavirus research in the COVID-19 crisis. We assessed the highest frequency topics, and the topic status (always alive, resurgent, or emerging), and research communities identified from the evolutionary pathways. Logistic regression

models were used to test the relationship between these selected topics and affiliation types, and geographical locations.

Following data extraction and manipulation (see Table 2), we retrieved 601,103 raw terms from combined titles and abstracts of 35,745 articles and identified 64,776 core terms on the coronavirus research by removing non-technical words and consolidating technical synonyms. In parallel, using the Word2Vec model, we collected 63,720 and 11,048 term vectors from articles published in pre-COVID-19 period and the COVID-19 period, respectively, which were then used as the input into the analysis of identifying knowledge clusters, described below.

Table 2 Stepwise term clumping process for identifying core terms on coronavirus-related research

Step	Description	#Terms
1	Raw terms retrieved by an NLP function integrated in VantagePoint	601,103
2	Remove meaningless terms, e.g., pronouns, prepositions, and conjunctions	594,116
3	Remove common terms in scientific articles, e.g., "methods"	584,465
4	Remove terms starting with non-alphabetic characters, e.g., "step 1" or "1.5 m/s"	517,502
5	Consolidate terms with specific rules, e.g., abbreviations and related full names	506,283
6	Remove terms appearing in only one record	89,497
7	Consolidate terms with the same stem, e.g., "infectious disease" and "infectious diseases"	81,871
8	Remove single-word terms, e.g., "virus"	68,055
9	Consolidate terms based on given topics, e.g., "MERS" and "MERS-COV"	64,776

4. Results

Identifying knowledge clusters

In a preliminary assessment of the types of research taking place during COVID-19 (Fry et al., 2020), we analyzed clusters of knowledge in coronavirus research before and during the crisis. To do so, we grouped similar terms in the two periods, respectively, and present the clusters of terms found in Table 3. The table shows the clustering of terms collected and illustrated in Figs. 2 and 3. Visuals from Fry et al. (2020) are reproduced here in Fig. 2 and Fig. 3. Specifically, Fig. 2 illustrates the topics derived from the articles produced by the coronavirus research community in the two years prior to the COVID-19 pandemic. We interpret this graph as exhibiting a well-ordered system of coronavirus research, which includes clusters of research surrounding SARS-CoV (Severe Acute Respiratory Syndrome Coronavirus) and MERS-CoV (Middle East Respiratory Syndrome Coronavirus) (two previous coronavirus outbreaks), which also happen to be the most common and most central topics. Other organizing topics are phylogenetic analysis, epidemiology, respiratory viruses, viral infection, and porcine epidemic diarrhea virus. Fig. 3 presents the topics in the first four months of the COVID-19 period where we see a more diverse and 'chaotic' set of research clusters around four broad topics: Wuhan, epidemiology, SARS-CoV, and fever. We suggest that Fig. 3 shows a knowledge ecosystem thrown into chaos by the pandemic and the scramble to gain information about what was occurring. Such observations lead to our key interests in understanding topic evolution, disruption, and resilience in early COVID-19 research from an ecosystem point of view. Fig. 4 shows the same community after nine months of research (May-October 2020) of COVID-19 and other associated coronavirus research.

The clear boundaries between clusters of terms, and cohesion within clusters in terms of the similarity in the knowledge base on coronavirus research in the pre-COVID-19 period shown in Table 3 coincide with our observation of ordered groups in Fig. 2. The pre-COVID-19 period shows distinct clusters of terms, such as epidemiology-related terms, virus-related terms, and clusters related to prior large coronavirus outbreaks ("SARS CoV" and "MERS CoV"). In contrast, Fig. 3 shows the eight clusters of terms identified in the initial shock of the pandemic which reveals a more chaotic situation. Fig. 3 shows the largest cluster to be 'Wuhan'-related terms, followed by COVID-19-related terms, which may reflect efforts to simply define the event. "SARS CoV," "MERS CoV" and "epidemiology" are retained from the pre-COVID-19 dataset, representing core pillars from the previous period. Fig. 4 shows the coronavirus research community after nine months of research—this figure represents articles published from the May-October 2020.

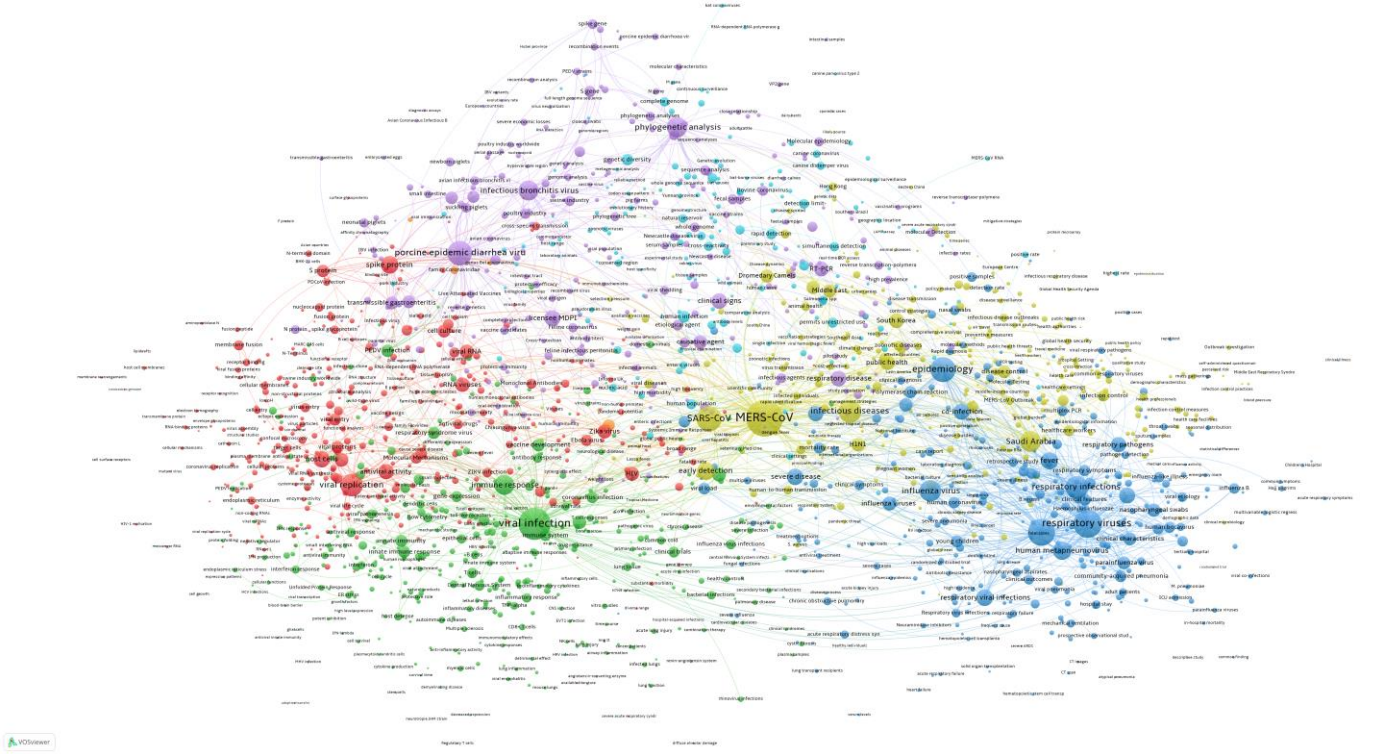

Fig. 2 Co-term map for the coronavirus research between 2018 and 2019
Note that this version was re-generated based on the data in the source: Fry et al. (2020)

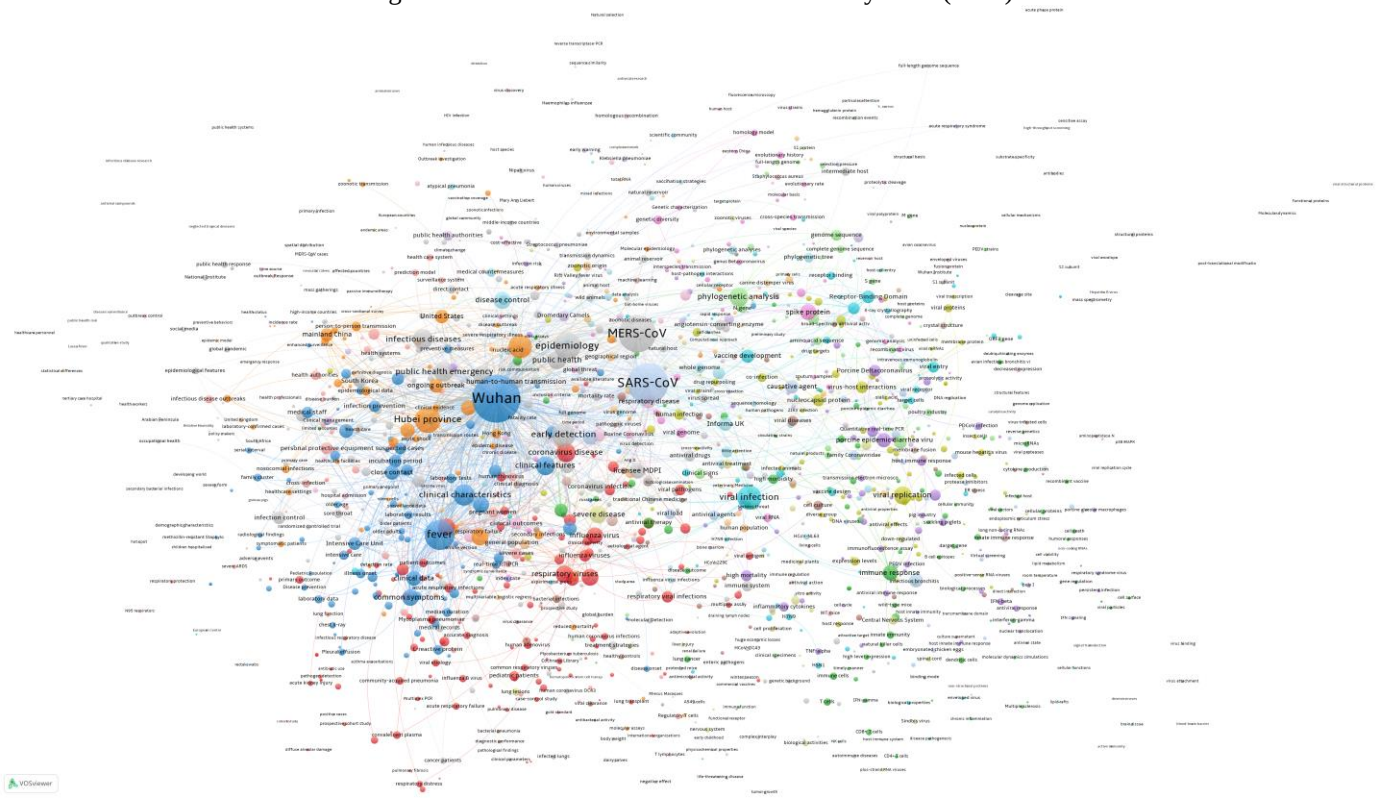

Fig. 3 Co-term map for COVID-19 research in early 2020 (January-April 2020)
Note that this version was re-generated based on the data in the source: Fry et al. (2020)

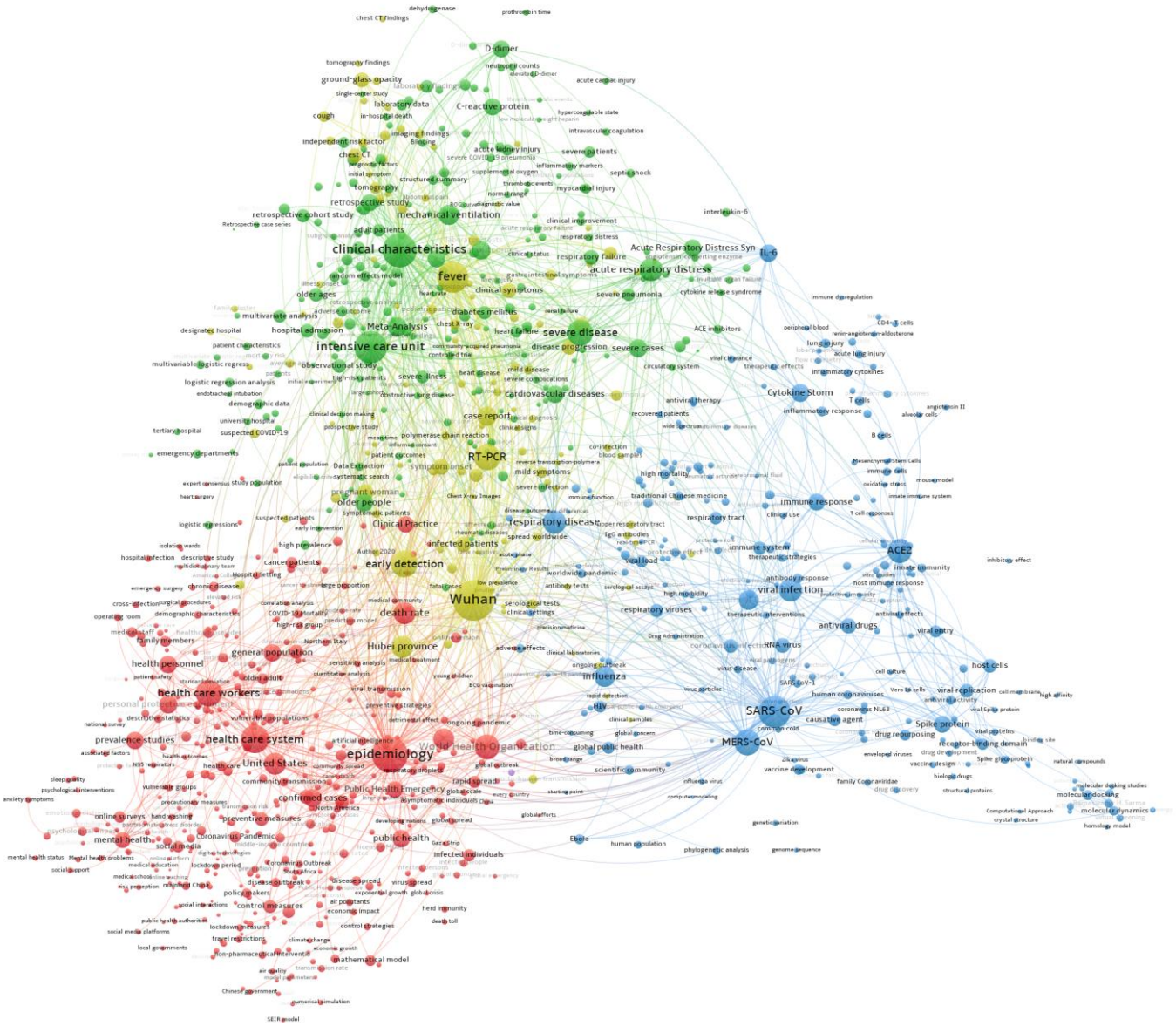

Fig. 4 Co-term map for the COVID-19 research in 2020 (May-October 2020)
Note: In order to assess medium to longer-term trends, publications in May-October 2020 were used.

Examination of the topic clusters shows that the epidemiology and immunology communities have been highly resilient and have reorganized and reemerged as research communities early in the pandemic. The genetics research community is nearly completely focused on Angiotensin-converting enzyme 2 (ACE2), which is an enzyme that attaches to the cell membranes--a feature of coronavirus infection. Next to immunology and epidemiology we see that patient care has persisted during the pandemic as major topic clusters (clinical characteristics, intensive care, respiratory distress, severe disease) while it was not apparent in pre-COVID research. Moreover, there remains a focus, although not as prominent as in the first days, on the geographic locations of the apparent locus of COVID-19 in Wuhan and Hubei Province, also not evident in the pre-COVID years. These two aspects—patient care and geographic focus--are completely new to the community as they did not pre-exist the COVID-19 pandemic. Some specific coronavirus diseases that were being researched prior to COVID-19 disappear from the map, as might be expected, while most of the community turns to the crisis. SARS and MERS both continue to appear in the clusters, however.

Measuring the centrality of the topic cluster networks shows that prior to the COVID pandemic coronavirus topics were highly decentralized with a betweenness centrality measure of 0.079 in the 10 years leading up to the pandemic. This suggests a broad frontier of research with multiple foci for research. Very early in the pandemic, the topic clustering becomes much more centralized with a betweenness centrality measure of 0.110. We can see this illustrated in Fig. 2 where the topics becomes hyper-centralized around ‘fever’ and ‘Wuhan’ and many disciplinary terms are greatly reduced. Search appears be highly constrained by symptoms and geography. Centrality drops in the May-October 2020 cluster, with the centrality measure of 0.013 now below the pre-COVID-19 period, suggesting a great deal of search and exploration with little focus on a frontier. An entire new cluster around patient care has been created.

Table 3 Topic extraction for the pre- and COVID-19 periods

ID	Topic label	Topic description
Pre COVID-19 period (2009-2019)		
1	epidemiology (996)	host cells (539), United States (463), infected cells (450), spike protein (393), co-infection (357), Central Nervous System (287), influenza-like illness (252), early stage (251), T cells (243), healthcare workers (228), antibody response (225), S protein (217), host response (203), nucleic acid (200), dendritic cells (190), nucleocapsid protein (181), Receptor-Binding Domain (176), cross-sectional study (175), flow cytometry (173), mammalian cells (172)
2	viral infection (1482)	viral replication (714), Saudi Arabia (585), public health (569), viral pathogens (371), viral RNA (368), viral proteins (320), World Health Organization (267), viral genome (264), human health (256), infection control (254), viral load (237), viral entry (228), genetic diversity (220), human infection (195), Intensive Care Unit (191), case report (184), interferon (177), viral diseases (173), PEDV infection (171), health care workers (168)
3	infectious diseases (1392)	fever (480), severe disease (417), cell culture (416), disease control (416), clinical signs (408), infectious agents (259), young children (248), feline infectious peritonitis (241), clinical trials (231), Dromedary Camels (219), clinical features (203), control group (199), developing countries (197), human disease (196), West Africa (193), clinical characteristics (189), Clinical presentation (179), IFN-gamma (159), prevention (155), common cold (152)
4	respiratory viruses (1081)	respiratory syncytial virus (1061), respiratory infections (462), respiratory viral infections (363), respiratory disease (358), respiratory tract infections (354), Middle East (317), acute respiratory infections (300), respiratory syndrome virus (300), respiratory tract (275), respiratory pathogens (249), acute respiratory distress syndrome (210), Feline coronavirus (203), respiratory virus (203), respiratory symptoms (201), coronavirus infection (182), Bovine Coronavirus (180), respiratory illness (164), human coronaviruses (161), human coronavirus (160), Acute respiratory tract infections (147)
5	SARS-CoV (2370)	immune response (771), HIV (588), gene expression (321), immune system (321), innate immune response (303), H5N1 (283), South Korea (280), Molecular Mechanisms (270), Monoclonal Antibodies (268), mouse model (256), study period (212), electron microscopy (202), inflammatory response (200), inhibitory effect (183), host immune response (175), molecular characterization (172), adaptive immune responses (166), mathematical model (156), endoplasmic reticulum (150), multiplex PCR (150)
6	MERS-CoV (2403)	phylogenetic analysis (683), antiviral activity (508), human metapneumovirus (408), animal models (352), causative agent (324), Hong Kong (301), real-time PCR (282), vaccine development (263), ages (233), human population (230), clinical samples (202), crystal structure (201), high mortality (200), age groups (198), human bocavirus (197), licensee MDPI (193), virus-host interactions (189), antiviral effects (180), fecal samples (179), etiological agent (177), mortality rate (177)
7	H1N1 (558)	RT-PCR (432), innate immunity (302), Polymerase chain reaction (273), sequence analysis (228), community-acquired pneumonia (197), rapid detection (193), Escherichia coli (187), enzyme-linked immunosorbent assay (182), H7N9 (180), cross-reactivity (173), real-time RT-PCR (171), complete genome sequence (168), complete genome (167), porcine epidemic diarrhea (160), Multiple sclerosis (150), nasopharyngeal aspirates (150), host factors (147), control measures (141), protective immunity (140)

8	porcine epidemic diarrhea virus (719)	infectious bronchitis virus (654), influenza virus (639), virus infection (552), RNA viruses (437), virus replication (421), influenza viruses (379), pandemic influenza (309), Ebola virus (271), transmissible gastroenteritis virus (268), mouse hepatitis virus (239), hepatitis C virus (229), avian influenza (211), virus entry (211), Dengue virus (186), Zika virus (186), enveloped viruses (176), influenza virus infections (161), Ebola virus disease (155), virus detection (152), influenza vaccination (149)
COVID-19 Period (2020)		
1	COVID-19 (2235)	COVID-19 outbreak (230), COVID-19 epidemic (127), clinical characteristics (116), United States (75), clinical features (74), mainland China (52), retrospective study (33), clinical manifestations (32), COVID-19 transmission (23), clinical outcomes (22), severe COVID-19 (22), clinical symptoms (21), Hong Kong (20), COVID-19 spread (18), traditional Chinese medicine (16), travel restrictions (16), Chinese government (15), retrospective cohort study (14), modeling studies (13), Case Study (12)
2	SARS-CoV-2 (751)	disease control (41), healthcare workers (34), common symptoms (27), chest CT (26), Saudi Arabia (24), viral pneumonia (24), Intensive Care Unit (23), CT images (21), Informa UK (21), global spread (20), clinical course (19), clinical practice (18), etiological agent (17), Molecular Mechanisms (17), SARS-CoV-2 outbreak (17), intensive care (16), SARS-CoV-2 pandemic (16), C-reactive protein (14), CT findings (14), viral genome (14)
3	Wuhan (635)	Hubei province (131), fever (99), coronavirus disease (62), confirmed cases (54), mathematical model (50), severe disease (49), coronavirus (41), epidemiological characteristics (39), spike protein (39), phylogenetic analysis (38), immune response (30), personal protective equipment (29), angiotensin-converting enzyme 2 (27), rapid spread (26), porcine epidemic diarrhea virus (21), retrospective analysis (21), severe pneumonia (21), suspected cases (21), severe cases (20), transmission dynamics (20)
4	SARS-CoV (254)	South Korea (40), incubation period (32), respiratory infections (31), early detection (24), cardiovascular diseases (21), preventive measures (15), Open Access article (14), co-infection (13), online version (13), viral load (13), high morbidity (12), exponential growth (11), cross-infection (10), Pleural effusion (10), acute respiratory infections (8), bacterial infections (8), Chinese General Practice (8), early identification (8), Feline coronavirus (8), medical countermeasures (8)
5	COVID-19 pandemic (237)	global pandemic (53), International Concern (51), ongoing outbreak (41), close contact (35), medical staff (33), causative agent (32), median age (30), imported case (24), Coronavirus Pandemic (23), coronavirus Outbreak (22), machine learning (20), healthcare systems (18), mechanical ventilation (17), global concern (14), case definition (13), Monoclonal Antibodies (13), real time (13), age groups (12), illness onset (12), diagnostic tests (11)
6	MERS-CoV (183)	early stage (43), case fatality rate (30), respiratory syncytial virus (28), early phase (26), host cells (23), Receptor-Binding Domain (22), mortality rate (21), respiratory illness (20), Cytokine Storm (19), infectious bronchitis virus (17), Shanghai Shangyixun Cultural Communication Co. Ltd (17), genome sequence (14), convalescent plasma (13), decision-making (12), intermediate host (12), adverse effects (11), family Coronaviridae (11), family members (11), John Wiley (11), serial interval (11)
7	epidemiology (112)	infectious diseases (91), ill patients (36), case report (35), urgent need (35), infected patients (31), clinical trials (30), general population (25), influenza virus (25), Clinical presentation (18), immune system (18), cancer patients (15), infected individuals (14), Clinical management (13), influenza viruses (12), Lopinavir/ritonavir (12), severe illness (12), antibody response (11), HIV (11), Northern Italy (11), pediatric patients (11)

8	World Health Organization (90)	public health (73), public health emergency (64), viral infection (56), control measures (40), acute respiratory distress syndrome (38), clinical data (33), infection control (30), pregnant women (30), respiratory viruses (30), coronavirus infection (29), global health (28), human-to-human transmission (28), respiratory disease (28), RT-PCR (27), viral replication (27), antiviral activity (26), vaccine development (25), licensee MDPI (24), symptom onset (23), infection prevention (22)

Note: The number following each term indicates the frequency of the term in the given dataset.

Evolutionary pathways

The raw dataset was run through the refined algorithm of Scientific Evolutionary Pathways (SEP) developed by Zhang et al. (2017b). This process produced 135 topics and 7 communities, with the predecessor-descendant relationships between these topics, which are plotted in an evolutionary pathway in Fig. 5. Table 4 shows the descriptive statistics for the basic results of the topic analysis including the numbers of records and terms.

Table 4 Descriptive statistics for SEP topics

		Max	Min	Average	Std. Dev.
Node	Number of terms	9483	1	237.23	867.33
	Number of articles	4837	1	457.53	704.01
Edge	Weight	0.1272	0.0003	0.0142	0.0162

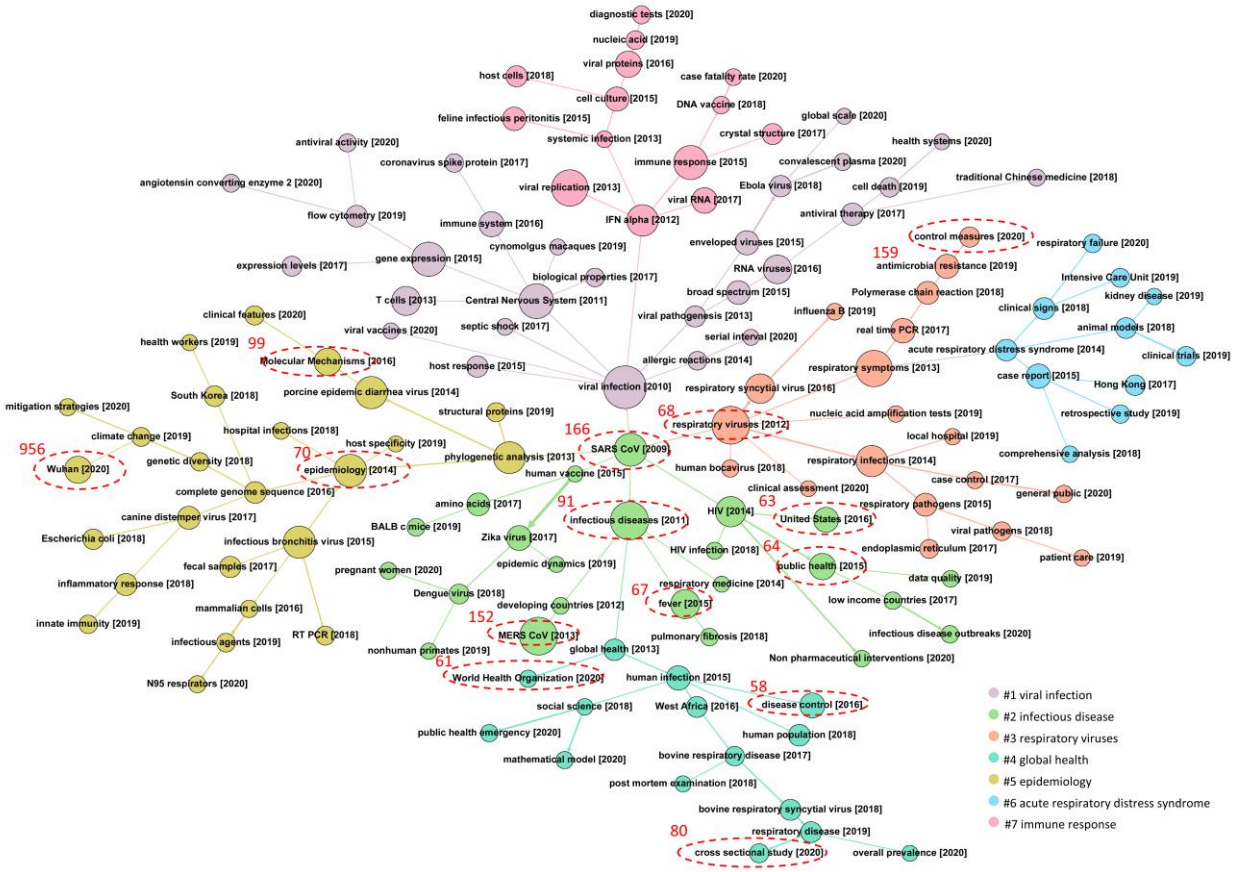

Fig. 5 Evolutionary pathways of the coronavirus research from 2009 to 2020
Note: Red dash circles mark topics where articles published/uploaded in 2020 are assigned, and the red digits indicate the number of those articles.

(1) Disruption and Resilience in the COVID-19 Crisis

Fig. 5 shows the evolutionary pathways for the full dataset of pre- and COVID-19 topic evolution. Examining the map using compass points, we defined "SARS-CoV" as the starting point in 2009 and it serves as the central point in the entire map with links leading in all directions. From SARS-CoV we see the evolutionary pathway spin off several lines of research mostly via the topics "viral infections" and "infectious diseases". We also see evolutionary pathways heading east into the community of topics under the header "respiratory

viruses." Looking northwest, from "viral infections," a line of research evolves into "central nervous system," which seeds genetic research activities (north). It is also worth pointing out that the new terms cooccur with the time of certain global or domestic epidemic outbreaks, such as "MERS CoV" (2013), "HIV" (2014), and "Zika Virus" (2017) although we do not see the name of cities or regions associated with those diseases as we see with Wuhan. This indicates a timely reaction conducted by the research community as a response to the outbreaks (Zhang et al., 2020; Porter et al., 2020).

Evolutionary pathways span out over time and newly born topics in 2020 could be observed in each community, indicating the disruption of those communities with the involvement of new knowledge in diverse levels. Among them, communities 1 (viral infection) and 4 (global health) have the largest number of new topics and community 1 appears to be more disrupted than other communities such as communities 6 (acute respiratory distress syndrome) and 7 (immune response), which might be considered relatively resilient in this COVID-19 crisis.

When tracing the assignment of those articles published/uploaded in 2020 (see red dash circles in Fig. 5), it is intriguing to see that two of the new topics emerging are "Wuhan" and "control measures." Similar to our earlier discussion, we interpret this as attempts to define the event given very limited knowledge. Having said that, as expected, in 2020 most pathways in the map return to the 'core pillars' of coronavirus research such as "infectious diseases" (including "SARS CoV" and "MERS CoV"), and "respiratory viruses"—the dominant species--whose knowledge bases have been well established for years. Similarly, along the pathway that was spawned by "phylogenetic analysis", we see "epidemiology" as a 'core pillar' for 2020 articles, along with topics "molecular mechanisms" and "Wuhan". Along the newly developed "global health" pathway, in addition to "disease control," we see two topics of note in 2020: "World Health Organization" and "cross-sectional analysis". Other pathways from viral infections are not a focus in the early days of COVID-19.

(2) Topic Similarity

In order to assess the level of disruption to the community during COVID-19 at the topic level, Table 5 shows topic similarity between all topics in the full sample by averaging the sum of cosine similarity between each focal topic and all other topics within a specified sample. Three broad similarity measures were created: 1) similarity between topics in the pre-COVID-19 period; 2) similarity between topics in the COVID-19 crisis, and 3) the similarity between the set of topics in the pre-COVID-19 period and the set of topics in the COVID-19 crisis. We find internal consistency of topics within the pre-COVID-19 period is 0.0432, and that of topics in the COVID-19 crisis is 0.0402. However, the consistency between the two sets of topics in pre- and during COVID is much lower, at 0.0267, which indicates different knowledge bases from one period to the other, or, a reorganization of the knowledge system around new priorities.

Table 5 Similarities of 2020 topics with topics in the pre-COVID-19 period

	Topic label	Similarity	Community
1	viral vaccines [2020]	0.0001	#1 viral infection
2	clinical assessment [2020]	0.0005	#3 respiratory viruses
3	serial interval [2020]	0.0015	#1 viral infection
4	global scale [2020]	0.0019	#1 viral infection
5	overall prevalence [2020]	0.0022	#4 global health
6	health systems [2020]	0.0091	#1 viral infection
7	pregnant women [2020]	0.0188	#2 infectious diseases
8	case fatality rate [2020]	0.0200	#7 immune response
9	Non pharmaceutical interventions [2020]	0.0216	#2 infectious diseases
10	public health emergency [2020]	0.0246	#4 global health
11	mathematical model [2020]	0.0247	#4 global health
12	convalescent plasma [2020]	0.0259	#1 viral infection
13	N95 respirators [2020]	0.0263	#5 epidemiology
14	World Health Organization [2020]	0.0278	#4 global health
15	mitigation strategies [2020]	0.0294	#5 epidemiology

	Topic label	Similarity	Community
16	angiotensin converting enzyme 2 [2020]	0.0305	#1 viral infection
17	general public [2020]	0.0306	#3 respiratory viruses
18	infectious disease outbreaks [2020]	0.0348	#2 infectious diseases
19	antiviral activity [2020]	0.0387	#1 viral infection
20	diagnostic tests [2020]	0.0404	#7 immune response
21	respiratory failure [2020]	0.0434	#6 acute respiratory distress syndrome
22	clinical features [2020]	0.0459	#5 epidemiology
23	cross sectional study [2020]	0.0508	#4 global health
24	control measures [2020]	0.0524	#3 respiratory viruses
25	Wuhan [2020]	0.0653	#5 epidemiology

From the perspective of community disruption and resilience, the disruption of community 1 "viral infection" shows that all its seven new topics in 2020 share low similarities with pre-COVID-19 topics. We interpret this to mean that "viral infection" cluster is disrupted and not resilient. In contrast, communities 5 "epidemiology" and 3 "respiratory viruses" appear to be more resilient with more terms coexisting in both the pre- and COVID-19 periods, from which the two largest newly born (emerging) topics in 2020 share the highest similarities with pre-COVID existing knowledge bases.

Further insights can be gained by examining topics that persist from the pre-COVID-19 period into the pandemic period, which ones die off, and which are newly introduced into the community in the pandemic. We defined those persistent topics as "always alive", those which resurge from earlier times as "resurgent", and those that appear for the first time in 2020 as "emerging". We identified 27 "always alive" topics, 9 "resurgent" topics (Table 6), as well as quite large number of 25 "emerging" topics (Table 5), suggesting significant disruption. Briefly, those "always alive" topics serve as the core pillars of the coronavirus research, "resurgent" topics might indicate specific interests raised along with technological change in the past decades, while "emerging" topics could represent frontier ideas or novel recombinations of past knowledge.

Table 6 reveals that the previous pandemics, namely "SARS CoV" and "MERS CoV" are persistent topics, representing stable pillars of the coronavirus research space. In contrast, most of the "resurgent" topics relate to common but distinctive symptoms of coronavirus infection, such as "fever" and "respiratory symptoms". It is possible that between pandemics these topics were not in frequent use amongst the research community but are needed once again to understand COVID-19. The topic "global health" is also a "resurgent" topic, which may indicate the urgent need for public health during the COVID-19 crisis, and perhaps an underinvestment in this capacity. As for other "emerging" topics, we see a range of technical topics, global health-focused topics, and ones topics related to clinical information and patient care. This diversity could represent the willingness of researchers to rapidly share hands-on experience with the virus—which may not have been published in years when preprint servers were not available. In the remainder of the paper we investigate these different types of topics using statistical analysis to better understand the specific interests of international collaborative communities and diverse affiliations.

Table 6 Status of sample topics

No	Label	Status	TF-IDF
1	Central Nervous System [2011]	Resurgent	0.5748
2	IFN alpha [2012]	Resurgent	0.4717
3	phylogenetic analysis [2013]	Resurgent	0.4851
4	respiratory symptoms [2013]	Resurgent	0.6230
5	viral replication [2013]	Resurgent	0.6211
6	global health [2013]	Resurgent	0.1675
7	acute respiratory distress syndrome [2014]	Resurgent	0.1721
8	cell culture [2015]	Resurgent	0.2483
9	fever [2015]	Resurgent	0.3979
10	SARS CoV [2009]	Always alive	0.5173
11	viral infection [2010]	Always alive	0.8175

No	Label	Status	TF-IDF
12	infectious diseases [2011]	Always alive	0.6786
13	respiratory viruses [2012]	Always alive	0.6649
14	MERS CoV [2013]	Always alive	0.6681
15	porcine epidemic diarrhea virus [2014]	Always alive	0.4979
16	epidemiology [2014]	Always alive	0.5101
17	infectious bronchitis virus [2015]	Always alive	0.4938
18	feline infectious peritonitis [2015]	Always alive	0.2278
19	immune response [2015]	Always alive	0.5533
20	public health [2015]	Always alive	0.3173
21	host response [2015]	Always alive	0.1614
22	respiratory pathogens [2015]	Always alive	0.2139
23	RNA viruses [2016]	Always alive	0.3874
24	viral proteins [2016]	Always alive	0.3138
25	respiratory syncytial virus [2016]	Always alive	0.4118
26	disease control [2016]	Always alive	0.2762
27	United States [2016]	Always alive	0.2718
28	viral RNA [2017]	Always alive	0.2913
29	fecal samples [2017]	Always alive	0.1640
30	crystal structure [2017]	Always alive	0.1541
31	Hong Kong [2017]	Always alive	0.1392
32	coronavirus spike protein [2017]	Always alive	0.0786
33	endoplasmic reticulum [2017]	Always alive	0.0888
34	amino acids [2017]	Always alive	0.2088
35	septic shock [2017]	Always alive	0.0276
36	biological properties [2017]	Always alive	0.0805

Statistical analysis of COVID-19 topics

(1) Topics and author location

Given the documented importance and benefits of international collaboration for scientific progress (Fry et al., 2020; Wagner et al., 2017), we further explored the relationship between international team structure of articles and article topics during the pandemic. Table 7 shows how different team structures (i.e., international, Chinese authorship, US authorship, China-US collaboration) correspond with the selection of topics during the COVID-19 pandemic. We find that international collaborative articles and those from the United States favor "always alive" topics, compared to resurgent or emerging topics. This finding supports other research that shows the tendency of international collaborative research towards conventional rather than novel research (Wagner et al., 2019). In contrast to international collaborative and US work, Chinese-authored articles are more likely to work on emerging topics as compared to "always alive" or "resurgent" topics. It may also be that, in the early days, Chinese researchers were facing unknown situations and needed to improvise their work more quickly than other regions.

To complement the findings from the statistical analysis on Chinese researchers, we focused specifically on the topics pursued by Chinese researchers compared to researchers from the rest of the world. Specifically, we re-ran the analysis of the evolutionary pathways to illustrate the pathways of articles emanating from China. To visually represent this, we adjusted the node size of the original SEP to represent the relative use of a given topic by Chinese based researchers in Fig. 6.

Fig. 6 reveals that compared to those dominant species topics (e.g., core pillars of coronavirus research), Chinese researchers tended towards emerging topics, with the largest nodes representing topics at the end of the evolutionary pathways and born in more recent years. Chinese research also puts more emphasis on selected communities, namely, community 1 "virus infection" and community 5 "epidemiology" - the most disruptive and resilient communities observed from Fig. 5. On the other hand, nodes in more stable

communities of research, community 7 “immune response” and community 4 “global health” are much smaller, indicating less frequently researched topics of Chinese researchers. This observed trend could be for a number of reasons, including the fact that Chinese researchers dominated early research on COVID-19 (Fry et al 2020). This early response to the outbreak, which corresponded to the location of the earliest cases, could present more opportunity to Chinese researchers to pursue novel research trajectories.

Table 7 Logistic regression on the relationship of authorship and topic status in COVID-19

Location of authors	Always alive					Resurgent					Emerging				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
International collaboration	0.262***					-0.335**					-0.167**				
	(0.084)					(0.167)					(0.084)				
Chinese authorship		-0.424***			-0.368***		0.153			0.148		0.366***			0.307***
		(0.079)			(0.092)		(0.140)			(0.162)		(0.078)			(0.090)
US authorship			0.402***		0.280***			-0.181		-0.098			-0.345***		-0.252**
			(0.086)		(0.103)			(0.167)		(0.207)			(0.086)		(0.104)
China-US collaboration				0.061	0.075				-0.153	-0.160				-0.020	-0.015
				(0.163)	(0.201)				(0.318)	(0.387)				(0.161)	(0.198)
Mean of the dependent variable	0.173	0.173	0.173	0.173	0.173	0.519	0.519	0.519	0.519	0.519	0.318	0.318	0.318	0.318	0.318
Obs.	2949	2949	2949	2949	2949	2949	2949	2949	2949	2949	2949	2949	2949	2949	2949
Pseudo R2	0.026	0.025	0.025	0.024	0.025	0.026	0.025	0.025	0.024	0.025	0.034	0.039	0.037	0.033	0.041

Note: Estimates stem from logistic regression models with dependent variables being dummy variables indicating status of topics (always alive, resurgent, emerging). Publication type and a dummy for whether the article is a preprint are controlled in all models.
Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

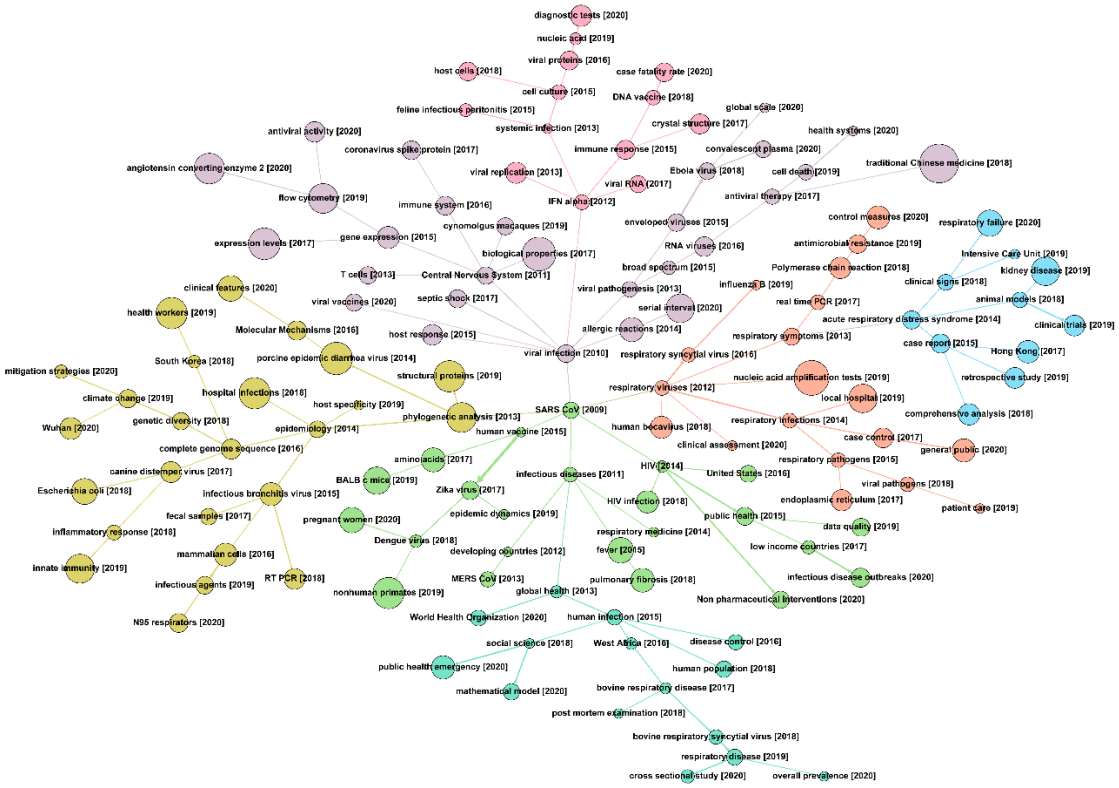

Fig. 6 Evolutionary pathways of the coronavirus research from 2009 to 2020 resized to show China's research emphases
Note: The size of nodes indicates the percentage of Chinese articles in global articles.

(2) Topics and author sector

We explored whether there is a relationship between authors with academic or industrial affiliations and the prevalence of always alive, resurgent, or emerging topics in Table 8. Research from government labs are more likely to use "always alive" topics and they are less likely to use resurgent topics. In contrast, academic or industrial researchers are more likely to use resurgent topics. Articles authored by academic researchers are also more likely to focus on emerging topics, although this difference is not statistically significant. These findings suggest that academic researchers, and industrial researchers (to a certain extent), are more likely to leverage unique recombination or pursue topics outside of the stable core pillars than those researchers affiliated with governmental organizations.

Table 8 Logistic regression on the relationship between researcher affiliations and topic status in COVID-19

Researcher type	Always alive				Resurgent				Emerging			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Academic	-0.047			0.010	0.037			-0.075	0.035			0.010
	(0.134)			(0.138)	(0.242)			(0.248)	(0.133)			(0.136)
Industry		0.538		0.562		-0.487		-0.540		-0.416		-0.423
		(0.357)		(0.358)		(0.731)		(0.721)		(0.367)		(0.367)
Government			0.285*	0.297*			-0.676*	-0.705*			-0.111	-0.116
			(0.164)	(0.168)			(0.370)	(0.382)			(0.165)	(0.169)
Mean of the dependent variable	0.384	0.384	0.384	0.384	0.074	0.074	0.074	0.074	0.541	0.541	0.541	0.541
Obs.	2949	2949	2949	2949	2949	2949	2949	2949	2949	2949	2949	2949
Pseudo R2	0.021	0.021	0.021	0.022	0.024	0.024	0.026	0.027	0.033	0.033	0.033	0.033

Note: Estimates stem from logistic regression models with dependent variables being dummy variables indicating status of topics (always alive, resurgent, emerging). Publication type and a dummy for whether the article is a preprint are controlled in all models.
Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

(3) Variation between communities

Fig. 7 shows the prevalence of topic types "always alive," "resurgent", and "emerging" in each topic community in COVID-19 articles. Overall, more than half of COVID-19 articles featured emerging topics. However, some communities have a larger proportion of emerging topics than others, such as communities 5 "epidemiology" and 4 "global health" (82% and 67%, respectively). It is necessary to point out that the largest proportion of 2020 articles in community 5 "epidemiology" is within the topic "Wuhan", which shares the highest similarity with epidemiology as generated before the COVID-19 crisis. In contrast, communities 2 "infectious diseases" and 7 "immune response" are much less likely to have emerging topics. This descriptive evidence is intriguing, but it is outside the scope of this paper to interpret the meaning of these facts to the coronavirus community. The reasons behind these variations is a subject for future research.

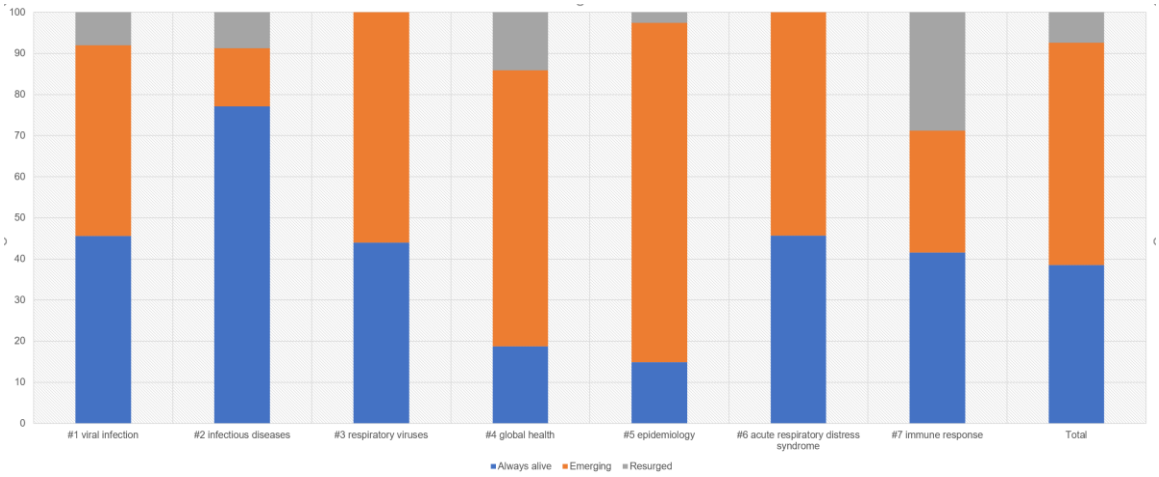

Fig. 7 Distribution of topic status in each community based on COVID-19 articles
Note: Among articles with each topic community in 2020, percentages of articles associated with always alive, emerging, and resurged topics are calculated, respectively.

5. Discussion and Conclusions

The onset of COVID-19 greatly disrupted the ecosystem of coronavirus research subjects that had formed into ordered clusters in the 10 years leading up to the crisis. Prior to the pandemic, the ecosystem was represented by well-organized clusters around SARS/MERS CoVs, porcine bronchitis virus, respiratory virus, viral infection, epidemiology, and phylogenetic analysis. Following the initial shock associated with the onset of COVID-19, topics reorganized, but around a different set of pillars than before the pandemic. In the initial shock period from January-April 2020, we see research related to definition of the event such as "Wuhan" as the largest cluster of research. In addition, research topics in the wake of COVID-19 retreat to strongly focus on SARS-CoV and MERS-CoV; these are the topics around which the whole community organized. We interpret this finding to mean that the SARS virus (which preceded the MERS epidemic in time) is the initiating event for this line of research and an evolutionary pathway that became COVID-19 research as the novel coronavirus appeared in 2019. As time progressed into the COVID-19 crisis we see that two research topic clusters have reemerged in the ecosystem: immunology and epidemiology. We interpret these communities to have high resilience, while other communities appear to have dispersed or reorganized into emergent communities.

While the pandemic in 2020 caused a disruption to coronavirus research, rendering some lines of research more peripheral than others, the community exhibits some return to core pillars of research, as would be expected of an ecosystem in crisis. It may be that over a longer period some subjects that had been the focus of research in the pre-COVID-19 period will diminish as attention is turned to COVID-19. Future research will seek to explore how the knowledge base adapts and changes as a response to the pandemic, and which of the actors (academic, industry, government) take the lead in the stabilization process.

Limitations

Although we consider the findings from the analysis as providing insight into the theoretical framework, the study has limitations. First, the analysis of COVID-19 as presented is based on a dataset of coronavirus articles published or posted on preprint servers before the end of April 2020. It is entirely possible that once pre-prints are subjected to peer review and articles are published, the research would present a different story. In fact, data collected later in 2020 showed an explosive growth of COVID-19 articles with about 39,000 new articles between April and mid-July 2020 and 43,500 new articles from mid-July to the start of October (Cai et al. forthcoming), and we anticipate different dynamics than those revealed in this data. Having said that, the theoretical predictions and focus of this study of the early months of the pandemic focused on the response of the ecosystem, and we can see clear patterns in the actions of the community. We explored the immediate response from a sudden shock. Future work will explore longer term consequences, including any possible returns to equilibrium.

Second, although we tried to define the involvement of different types of researchers in terms of their affiliation, i.e., academic, industry, and government, we acknowledge that the classification of the affiliation types (through the use of affiliation text in articles) is limited. For example, the involvement of government institutions is likely to be an underestimate as some national labs that are affiliated with universities are counted as 'academic' rather than 'government'. In China in particular, national labs are always listed as a sub-institution of a university or an institute, therefore they do not contribute to the publication shares of 'government'.

In the SEP methodologies, labels are derived from scientific terms, based on frequency of occurrence in topics. We used the most representative term to label a topic, which in most cases is a high-frequency term, but duplicate terms are not allowed, even if subjects are closely aligned. This means that topics might result in the label that does not well represent the central ideas of involved articles, but in the next emergent term on the pathway. As we see with the term surge of the term "Wuhan" the topic itself is part of evolutionary emergence rather than representing underlying science. Thus, it is important to see the terms as the evolution of knowledge pathways rather than the advancement of science in individual topics.

Acknowledgments

Yi Zhang and Mengjia Wu are supported by the Australian Research Council under Discovery Early Career Researcher Award DE190100994. Xiaojing Cai would like to acknowledge the support from Fulbright Foreign Student Program and the National Natural Science Foundation of China (Grant number: 71843012).

Declarations

Conflicts of interests

All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Availability of data, material, and code

All data, materials, and code of this study are available on request to yi.zhang@uts.edu.au

Author contributions

Designed research: Yi Zhang, Xiaojing Cai, Caroline V. Fry, Caroline S. Wagner; Performed research: Yi Zhang, Xiaojing Cai, Caroline V. Fry, Mengjia Wu, Caroline S. Wagner; Contributed new reagents or analytic tools: Yi Zhang, Mengjia Wu; Analyzed data: Yi Zhang, Xiaojing Cai, Caroline V. Fry, Mengjia Wu; Wrote the paper: Yi Zhang, Xiaojing Cai, Caroline V. Fry, Caroline S. Wagner.

References

Acemoglu, D., & Linn, J. (2004). Market size in innovation: Theory and evidence from the pharmaceutical industry. *The Quarterly Journal of Economics, 119*(3), 1049-1090.

Allan, J. (2012). *Topic detection and tracking: Event-based information organization* (Vol. 12): Springer Science & Business Media.
Azoulay, P., Fons-Rosen, C., & Graff Zivin, J. S. (2019). Does science advance one funeral at a time? *American Economic Review, 109*(8), 2889-2920.
Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., et al. (2020). A large-scale COVID-19 Twitter chatter dataset for open scientific research--an international collaboration. *arXiv preprint arXiv:2004.03688*.
Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. Paper presented at the *Third International AAAI Conference on Weblogs and Social Media*, Menlo Park, California.
Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993-1022.
Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: The UCSD map of science. *PloS One, 7*(7), e39464.
Borrett, S. R., Moody, J., & Edelmann, A. (2014). The rise of network ecology: Maps of the topic diversity and scientific collaboration. *Ecological Modelling*, 293 (2014): 111-127.
Cai, X., Fry, C. V., & Wagner, C. S. (2021). International collaboration during the COVID-19 crisis: Autumn 2020 developments. *Scientometrics*. Doi: 10.1007/s11192-021-03873-7
Catalini, C., Fons-Rosen, C., & Gaulé, P. (2020). How do travel costs shape collaboration? *Management Science, 66*(8), 3340-3360.
Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology, 61*(7), 1386-1409.
Colavizza, G., Costas, R., Traag, V. A., van Eck, N. J., van Leeuwen, T., & Waltman, L. (2020). A scientometric overview of CORD-19. *PloS One, 16*(1), e0244839.
Contractor, N. S., Wasserman, S., & Faust, K. (2006). Testing multitheoretical, multilevel hypotheses about organizational networks: An analytic framework and empirical example. *Academy of Management Review, 31*(3), 681-703.
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.
Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. *Journal of the Association for Information Science and Technology, 65*(10), 2084-2097.
Finkelstein, A. (2004). Static and dynamic effects of health policy: Evidence from the vaccine industry. *The Quarterly Journal of Economics, 119*(2), 527-564.
Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117-132.
Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal, 25*(8-9), 909-928.
Folke, C. (2006). Resilience: The emergence of a perspective for social–ecological systems analyses. *Global Environmental Change, 16*(3), 253-267.
Fry, C. V., Cai, X., Zhang, Y., & Wagner, C. S. (2020). Consolidation in a crisis: Patterns of international collaboration in early COVID-19 research. *PLoS One, 15*(7), e0236307.
Furman, J. L., & Teodoridis, F. (2020). Automation, research technology, and researchers' trajectories: Evidence from computer science and electrical engineering. *Organization Science, 31*(2), 330-354.
Ganguli, I. (2015). Immigration and ideas: What did Russian scientists "bring" to the United States? *Journal of Labor Economics, 33*(S1), S257-S288.
Gao J, Barzel B, Barabási A (2016). Universal resilience patterns in complex networks. *Nature,* 530(7590), 307-312.
Holling, C. S. (1973). Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics, 4*(1), 1-23.
Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.
Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR), 31*(3), 264-323.

Kokudo, N., & Sugiyama, H. (2020). Call for international cooperation and collaboration to effectively tackle the COVID-19 pandemic. *Global Health & Medicine, 2*(2), 60-62.
Kyhlstedt, M., & Andersson, S. W. (2020). Diagnostic and digital solutions to address the COVID-19 pandemic: The need for international collaboration to close the gap. *Health Policy Technol, 9*(2), 126-128.
Lee, S., & Monge, P. (2011). The coevolution of multiplex communication networks in organizational communities. *Journal of Communication, 61*(4), 758-779.
Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation, 29*(6-7), 481-497.
Myers, Kyle, (2020), The elasticity of science, *American Economic Journal: Applied Economics*, *12*(4), 103-34.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv*:1310.4546.
Mohamed, K., Rodríguez-Román, E., Rahmani, F., Zhang, H., Ivanovska, M., Makka, S. A., et al. (2020). Borderless collaboration is needed for COVID-19—A disease that knows no borders. *Infection Control & Hospital Epidemiology*, 1-2.
Monge, P. R., Contractor, N. S., Peter, R., Contractor, P. S., & Noshir, S. (2003). *Theories of Communication Networks*. Oxford University Press, USA.
Monge, P., Heiss, B. M., & Margolin, D. B. (2008). Communication network evolution in organizational communities. *Communication Theory, 18*(4), 449-477.
Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), 8577-8582.
Porter, A. L., Zhang, Y., Huang, Y., & Wu, M. (2020). Tracking and Mining the COVID-19 Research Literature. *Frontiers in Research Metrics and Analytics, 5*, 12.
Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*: McGraw-Hill, Inc.
Simon, H. A. (1991). The architecture of complexity. In *Facets of Systems Science* (pp. 457-476): Springer.
Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450-1467.
Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464-2476.
Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics, 59*(3), 467-472.
Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy, 34*(10), 1608-1618.
Wagner, C. S., Whetsell, T. A., & Leydesdorff, L. (2017). Growth of international collaboration in science: revisiting six specialties. *Scientometrics, 110*(3), 1633-1652.
Wagner, C. S., Whetsell, T. A., & Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy, 48*(5), 1260-1270.
Walker, B., & Salt, D. (2012). *Resilience thinking: Sustaining ecosystems and people in a changing world*: Island press.
Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B, 86*(11), 471.
Zhang, L., Zhao, W., Sun, B., Huang, Y., & Glänzel, W. (2020). How scientific research reacts to international public health emergencies: a global analysis of response patterns. *Scientometrics, 124*, 747-773.
Zhang, Y., Chen, H., Lu, J., & Zhang, G. (2017a). Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016. *Knowledge-Based Systems, 133*, 255-268.
Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., et al. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099-1117.
Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26-39.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change, 105*, 179-191.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017b). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology, 68*(8), 1925-1939.