Convergence in Viral Outbreak Research: Using Natural Language Processing to Define Network Bridges in the Bench-Bedside-Population Paradigm

**Margaret V. Powers-Fletcher¹ Erin E. McCabe² Sally Luken^{3,4}
Danny Wu^{5,6} Philip A. Hagedorn^{7,5} Ezra Edgerton² Amy Koshoffer^{8,9}
Dorcas Washington¹⁰ Suraj Kannayagari¹¹ Jason Lee¹²
Jennifer Latessa¹³ Anita Shah^{14,7} James Jaehoon Lee^{15,16,11}**

¹Division of Infectious Diseases, Department of Internal medicine, College of Medicine, University of Cincinnati, Cincinnati, Ohio, United States of America,

²Digital Scholarship Center, University of Cincinnati, Cincinnati, Ohio, United States of America,

³Luken Solutions, Columbus, Ohio, United States of America,

⁴Department of English, College of Humanities and Fine Arts, University of Massachusetts Amherst, Amherst, Massachusetts, United States of America,

⁵Department of Biomedical Informatics, College of Medicine, University of Cincinnati, Cincinnati, Ohio, United States of America,

⁶Center for Health Informatics, College of Medicine, University of Cincinnati, Cincinnati, Ohio, United States of America,

⁷Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, Ohio, United States of America,

⁸Research and Data Services,

⁹University of Cincinnati Libraries, University of Cincinnati, Cincinnati, Ohio, United States of America,

¹⁰Health Science Library, University of Cincinnati Libraries, University of Cincinnati, Cincinnati, Ohio, United States of America,

¹¹University of Cincinnati, Cincinnati, Ohio, United States of America,

¹²Flatiron Health, New York City, New York, United States of America,

¹³Research and Data Services Team, Science and Engineering Library, University of Cincinnati Libraries, University of Cincinnati, Cincinnati, Ohio, United States of America,

¹⁴Division of Hospital Medicine, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, Ohio, United States of America,

¹⁵Digital Scholarship Center, College of Arts and Sciences, University of Cincinnati Libraries, University of Cincinnati, Cincinnati, Ohio, United States of America,

¹⁶Department of English and Comparative Literature

Published on: Mar 22, 2021

DOI: <https://doi.org/10.1162/99608f92.cc479d52>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Research on viral outbreaks at the pandemic scale responds to heightened social urgency and the need to expedite scientific discovery from the ‘bench’ to the ‘bedside’ to the wider population. We sought to better understand translational research within the context of pandemics, both historical and present day, by tracking publication trends in the immediate aftermath of virus outbreaks. We used a blend of natural language processing (NLP), social network analysis, and human annotation approaches to analyze the 85,663 articles in the COVID-19 Open Research Dataset (CORD-19). We found stable and repeated characteristics throughout subsets of peer-reviewed published literature corresponding to seven different viral outbreaks over the last several decades. Three distinct groups or ‘neighborhoods’ recurred across all of the model networks: (1) bench science, (2) clinical treatments, and (3) broader public health trends. Notably, in each historical virus model, small ‘bridge’ nodes representing translational research connect the three otherwise disconnected neighborhoods. These bridging studies embody research convergence by both integrating the vocabulary and methods of different disciplines and bodies of previous work and by citing other papers beyond their narrow field. In the case of COVID-19, the literature continues to evolve apace along with the virus, and we can witness the phases of response unfold as the science progresses. This study demonstrates how the different sectors of biomedical research respond independently to public health emergencies and how translational research can facilitate greater information synthesis and exchange between disciplinary silos.

Keywords: interdisciplinary research, team science, natural language processing, social network analysis, COVID-19, biomedical informatics

Media Summary

As the COVID-19 outbreak exploded in 2020, our team witnessed a fragmented and often confused response to the virus in the scientific discourse, as well as in popular and social media. Drawing inspiration from the expression that “history never repeats itself but it rhymes,” we investigated whether underlying patterns occurred in the clinical and basic science response to virus outbreaks of the past several decades. To accomplish these goals, we assembled a diverse team of clinicians, data scientists, and experts in the analysis of language data to see how the research community responded to previous virus outbreaks in the published scientific literature.

We used the CORD-19 data set (containing 85,663 articles to July 14, 2020) aggregating existing research studies on viruses from peer-reviewed journals indexed in PubMed, and also from preprint platforms like medRxiv and ChemRxiv. We used a wide range of human and machine-based methods, drawing from the diverse expertise of our team—from machine learning methods that identify the structure of language patterns across all of the articles, to social network analysis that visualize clusters and identify hotspots of linguistic

activity across articles, authors, and scientific disciplines. All of our machine-based methods had a foundation of human benchmarking, where human coders labeled and assessed our models to validate the machine-based steps with the critical eyes of domain experts.

Using this hybrid methodology, we created concept maps of the existing virus research from the past 50 years for coronaviruses (SARS-CoV, MERS-CoV, and SARS-CoV-2) and noncoronaviruses (HIV, Zika, H1N1, and Ebola). From these models, we were able to identify stable and repeated patterns in how the research community mobilized in the face of historical virus outbreaks. Three distinct groups or ‘neighborhoods’ recurred across all of the model networks— (1) foundational laboratory science, (2) clinical treatments, and (3) broader public health trends. Notably, in each historical virus model, small ‘bridge’ clusters of articles and concepts representing cross-disciplinary research connect the three otherwise disconnected neighborhoods. These bridging research studies embody knowledge synthesis by both integrating the vocabulary and methods of different disciplines and bodies of previous work and by citing other articles and authors beyond their narrow field. In the case of COVID-19, the scientific research continues to evolve apace along with the virus, and we describe how social media, politics, and the global scope of the pandemic has created slight but perceptible variations in the virus discourse. This study demonstrates how the different sectors of biomedical research respond separately to public health emergencies. Cross-disciplinary research teams have facilitated greater information synthesis and exchange between isolated pockets of scientific knowledge to accelerate the pace of discovery to successfully overcome outbreaks of epidemic and pandemic proportions.

1. Introduction: Viral Outbreak Literature and Natural Language Processing

Viral outbreaks that reach epidemic or pandemic scales generate a stable and repeated research response from the biomedical research community. This response is, to some degree, a reflection of the exigency of a pandemic’s social response. Working in a pandemic context, biomedical researchers pivot their research focus and accelerate or intensify their work in response to the more immediate needs of the broader population. In November 2019, Bedford et al. (2019, p. 131) wrote that the “need to understand major trends in research and how and when they may influence the response to an epidemic” was critical to our success in responding to an infectious disease outbreak. One month later, the first case of a disease caused by a novel coronavirus was reported in the Hubei Province of China (World Health Organization, 2020); 6 months later, the virus had spread, resulting in over 4.7 million cases worldwide of the Coronavirus Disease 2019 (COVID-19) and over 300,000 reported COVID-19 deaths worldwide.

Often, there are barriers that delay clinical research and may prevent it from being “fast, flexible and integrated with the frontline response” to an epidemic (Gobat et al., 2019). To overcome these barriers, members of the Global Research Collaboration for Infectious Disease Preparedness (GloPID-R) and the Platform for European

Preparedness Against (Re-)emerging Epidemics (PREPARE) suggest that clinical research must use innovative design and delivery of results, improve the environment in which research occurs, and, importantly, strive to create multidisciplinary partnerships and collaborations (Gobat et al., 2019). This multidisciplinary collaboration is particularly important for clinical and translational research efforts, which leverage the findings of basic, or ‘bench,’ science to directly affect patient care. When these findings are applied more broadly, they have the potential to affect population health as well. In this way, a pipeline is established to promote the advancement of scientific discovery from the ‘bench’ to the ‘bedside’ to the population.

The goal of this study is to better understand the framework of translational research within the context of viral outbreaks, both historical and present day. Through a combination of qualitative and mixed-methodology approaches that include text mining and natural language processing (NLP), we find stable and repeated characteristics throughout the subset of peer-reviewed published literature found within the COVID-19 Open Research Dataset (CORD-19). We assess the nature of this repeated pattern of research output through the use of topic models that allow us to consider the highest probability clusters of co-occurring word- and article-level relationships that exist in the scientific literature responding to these notable viral epidemics of the last several decades. Content experts then evaluate these topic clusters and corresponding bridge relationships to assess their clinical and scientific significance. This approach allows us to validate the findings of the machine learning algorithms and elevate the discourse beyond simple pattern recognition. In doing so, we gain valuable insight related to 1) how the trends that we observe across viral outbreaks may point to innovative research or guide policy development and 2) how the different sectors of medical research communicate with each other to synthesize and transmit information.

2. Methods: Team Science Combining Qualitative and Quantitative Expertise

2.1. Opening the Black Box Through Qualitative Information Retrieval of Underlying Data

Our multidisciplinary team brings together expertise in both qualitative and quantitative methods, and, accordingly, our goal is to put into practice a hybrid approach that aims to increase confidence in, and the replicability of, research on unstructured text data sources. Our primary contribution does not depend on the assumption that an NLP or pattern recognition algorithm will lead to more precise and replicable results from studies of scientific literature, which has been described as the “science of science” (Talley, 2011; Zeng, 2017).

For transparency and reproducibility of our results, we supplement our sequence of decisions and specific parameters described below with [the underlying code](#) and [derived data](#) in the Appendices. For internal validation of our models, we evaluated topic coherence (Section 2.3), and for external validation, we performed a randomized human annotation agreement test (Section 2.4).

For the sake of replicability, we deploy the pattern recognition capabilities of NLP methods as an information retrieval, and not a black-box classification, method to provide models capable of evaluation by our panel of multiple independent coders. A hybrid NLP approach using human judgment to verify and tag the machine-based result outperforms a purely machine-based analysis (Chang, 2009). We use NLP as a way to explore patterns in our corpus as an initial information-retrieval step that identifies major trends at the word and document levels in the scientific literature. However, the trained judgment of subject matter experts from infectious diseases and other clinical fields directly evaluated the language and documents underlying the models to shape our claims through a system of intercoder agreements that we required for each step of analysis.

Because Latent Dirichlet Allocation (LDA) produces both word- and document-level relationships within a body of texts, it can serve as an information-retrieval logic that delivers the most relevant word groupings and underlying documents to subject matter–expert human evaluators. Although LDA has been used for classification, our information-retrieval implementation hews closer to its original stated design as a text categorization technique (Boyd-Graber, 2017). Topic models produce reliable models when used as an information-retrieval approach that takes into account both word- and document-level results, but become more fickle when the document-level results are not used as a second layer of verification to support the word-level output (Chang, 2009).

Our use of NLP as an information-retrieval technique makes linguistic and textual patterns in the scientific literature more tractable to well-tested, mixed-methods approaches such as intercoder agreement and validation that can increase confidence and replicability of qualitative research. Our conclusions therefore should not be taken as the result of arbitrary or purely subjective interpretations of models stripped from context. Nor do we place our faith blindly in the black box of any given machine learning algorithm. Our hybrid qualitative / quantitative methodology reflects the diversity of our team, and correspondingly the human team members articulate and test claims and connect the dots of inference based on our model of models technique as a machine-based information retrieval system.

2.2. Data Set Preprocessing and Parameter Selection

We collected coronavirus-related scientific literature using the CORD-19 data set released by the Allen Institute of Artificial Intelligence as of July 14, 2020 (Wang et al., 2020). We focused on the articles with full text and removed the non-English articles, duplicates, and preliminary papers; this resulted in 85,663 total articles for analysis. We then retrieved the full text of the articles and parsed them to extract paragraphs and strings related to publication (e.g., DOI numbers) to eliminate potential false-positive word relationships created by ubiquitous bibliographical metadata features unrelated to our analysis. In addition to cleaning and filtering article text, we also collected the metadata of these articles from the CORD-19 metadata data set and cleaned the dates into a uniform format for analysis.

For corpus preprocessing, we did not apply stemming or lemmatization. We employed the Allen Institute for AI's ScispaCy biomedical text processing package's tokenizer, which includes n-grams and part of speech tagging for scientific language. Our n-gram tokenization is based on ScispaCy's NLP library because of its specific tailoring for biomedical text analysis, specifically [the en_cor_sci_lg_model](#). We created a filtered dictionary for each subcorpus, keeping only words that occurred in fewer than 75% of documents or greater than five documents.

We also elected to include no stop words (apart from those included in the spaCy list) in any of the models. The domain specificity of ScispaCy's vocabulary obviated the need to use conventional English stop word lists. Using minimal stop words allowed us to observe the more often ignored, but not entirely ubiquitous, language for latent, meaningful patterns among more specified language. We aimed to depict not only where and how pandemic language separates itself in these models through difference, but also where and how near-ubiquitous, general language links these disparate areas.

Next, we indexed the full-text and the metadata of the article corpus using the Elasticsearch engine hosted by the Digital Scholarship Center (DSC). The DSC is part of the University of Cincinnati Libraries and has considerable experience in processing and analyzing scientific literatures in various domains. We applied this expertise to generate nine models: a full-corpus index model as a reference standard, a model based on 10,000 randomly selected documents from the corpus, and seven different virus-specific models corresponding to key viral outbreaks from the past several decades. The seven viruses selected for this analysis included coronaviruses (SARS-CoV, MERS-CoV, and SARS-CoV-2) and noncoronaviruses (HIV, Zika, H1N1, and Ebola) (Figure 1). Each of the nine models was run in six replicates for a total of 54 models. The 10,000-document randomized model produced a structure that reflected a generalized nonspecific virus model, which is unsurprising since the corpus is already preselected to the narrow range of COVID-19-related virus scientific literature. For a more accurate randomized comparison, we ran a model on 10,000 random articles from PubMed Central, which did not share the attributes of our COVID-19 virus corpus models.

To create these targeted subsets of the articles, we used two or more occurrences of virus-specific keywords drawn from conventions established by the World Health Organization, for example, SARS + “Severe Acute Respiratory Syndrome” (Table 1).

Table 1. Number of documents in subcorpora. (See also Figure 1.)

Search Terms	Number of Documents
Total corpus	85,663

"severe acute respiratory syndrome coronavirus 2" OR "covid-19" OR "sars-cov-2"	31,818 (2,192 from 01/01/2020 to 04/01/2020, 25,169 from 04/01/2020 to 07/01/2020)
SARS OR "Severe Acute Respiratory Syndrome"	30,383
MERS OR "Middle East Respiratory Syndrome"	8,855
HIV OR "human immunodeficiency virus" OR AIDS OR "acquired immunodeficiency syndrome"	12,024
H1N1 OR "Influenza A"	12,720
Zika	1,096
Ebola OR EVD OR "Ebola Virus Disease"	4,089

VIRUS ARTICLES (CORD-19 Dataset)

Figure 1. Number of documents in subcorpora.

As members of the same family of viruses, SARS-CoV, MERS-CoV and SARS-CoV-2 (SARS, MERS, COVID-19) share many features, including similar terminology from both a viral mechanistic standpoint and, often, a clinical perspective. By contrast, noncoronaviruses, such as HIV, Zika, H1N1, and Ebola, vary more

widely and fundamentally in virology, clinical characteristics, and terminology. By including both coronavirus and noncoronavirus outbreaks within our analysis, we aimed to not only observe the research response to large viral outbreaks generally, but also establish an external benchmark to which the COVID-19 and other coronavirus models could be compared. In turn, each virus model provides a sequential contribution to a historical perspective of viral epidemics over the past several decades of published literature.

Each set of virus-specific keywords was searched using the Elasticsearch engine to retrieve relevant documents. We then analyzed these subgroups of documents using visualizations arranged by the DSC's data mining "model of models" (MoM) platform (Lee & Beckelheimer, 2020). The DSC's platform is an implementation of LDA topic modeling, a machine learning algorithm used to observe the latent patterns, or 'hidden structures,' in large corpora of language data (Blei, 2012).

The application of LDA topic modeling to large corpora of scholarly literature, especially in scientific disciplines, has increased in recent years, with several teams exploring the latent language structures observable in scholarly literature through similar methods. Baghaei Lakeh and Ghaffarzadegan (2017), for example, employ LDA in their analysis of the behavioral and social science academic response to HIV/AIDS over three decades. Tshitoyan et al. (2019, p. 98) argue that "[s]uch language-based inference methods can become an entirely new field of research at the intersection between natural language processing and science." In this sense, LDA can provide a useful means of parsing large and ever-growing collections of scientific literature at a macro-level, which can, in turn, help scientists "choose the most promising way forward" (Tshitoyan et al., 2019).

In LDA, it is assumed 'that documents exhibit multiple topics' composed of words that display a high probability of co-occurring in a document or group of documents. So, for example, based on LDA we may infer that a topic composed of the words 'cat, dog, fish, pets, animals' has some latent relationship based on linguistic usage that differs from 'beef, chicken, pork, ham, meat.' Although both topics invoke animals, the usage of the words differs in the corpus. The LDA algorithm uses these co-occurrence patterns to produce a matrix of relationships between words across a corpus, and topics across all documents composing a corpus (Blei, 2012). In our application of LDA topic modeling, we consider the highest probability topic-clusters of word- and article-level relationships that exist in the full corpus and the seven virus-specific subsets of the 85,663 articles. We ran all models with identical input parameters in this study in both Gensim and Apache Spark to verify that our results were not the false positive artifact of quirks in either particular library. As a hierarchical Bayesian model, LDA uses prior distribution of latent topics governed by established alpha and beta hyperparameters (George & Doss, 2018). These alpha and beta hyperparameters both reflect concentration (aka density) in the prior distribution on a scale of 0 to 1. The alpha hyperparameter determines the assumed concentration of topics per document, and the beta hyperparameter determines the assumed concentration of words associated with a topic. As alpha approaches 1, more specific topics emerge while as beta approaches 1, more specific "word distribution[s] per topic" emerge (Kapadia, 2019). These hyperparameter values can be

adjusted according to the needs of any given corpus, usually based on the comparative coherence scores of several trials or human evaluation. As we describe in the following validation steps, we selected an alpha parameter value of 0.1 and a beta parameter of 0.01.

2.3. Internal Validation: Topic Coherence

Given the numerous models on different subsets of a single large corpus in our study, we chose to use model coherence scores as a measure for hyperparameter tuning rather than reassigning hyperparameters for every model. A 0.1/0.01 alpha/beta governance was chosen because this combination of parameters yielded the highest topic coherence scores over a 10,000-document random sample of our data.

Topic coherence leverages the *distributional hypothesis* of linguistics, that “words with similar meanings tend to occur in similar contexts” (Syed & Spruit, 2017, p. 2). It calculates this score based on whether the top 10 words in each topic are related in the context of the corpus. Several coherence measures have been shown to correlate with human interpretability (Mimno et al., 2011). In our case we used the context vector measure, C_v . Coherence is calculated through a four step pipeline: creating sets of pairs of words from the top words in each topic, calculating the relative frequency of each of those word pairs in the documents using a sliding window of 110 words as a ‘document,’ calculating confirmation measures showing how strongly the top word sets support one another using C_v , and taking the arithmetic mean of the topics’ coherence scores as the final model coherence score (Röder et al., 2015). For evaluation we ran models with a fixed random seed on the 10,000 random-document sample, enumerating through each of the following hyperparameter settings, resulting in 288 models:

Alpha = [.01, .1, .31, .91, asymmetric]

Beta = [.01, .1, .31, .91, auto]

Topic number = 20, 30, 40, 50, 60, 70, 80, 90

An alpha of 0.1 and a beta of 0.01 performed best on average, resulting in mean coherence scores of 0.53 and 0.54 across all other settings. Each of these settings accounted for five and seven, respectively, of the top 10 overall coherence scores. ([Supporting data is available here.](#)) We used these high-performing priors for our disease-specific models. Choosing the fixed priors based off average performance in general rather than the single highest performing set of hyperparameters was intended to create a more universal high-coherence model when training on a new subcorpus of the larger corpus.

Subsequently, given that the number of topics chosen significantly affects model interpretability, for each of the seven disease-specific models we ran coherence evaluation with the chosen alpha and beta priors and topic numbers ranging from 15 to 75 incremented by 5. Of these evaluated models, we chose the model run on that subcorpus with the number of topics that maximized coherence for our final analysis. These high-performing

topic numbers ranged from Zika at 15 topics (coherence .45) topics to HIV at 55 topics (coherence .58). ([Supporting data is available here.](#))

Our results are also governed by model-specific parameters that are variable on the DSC's platform. These included a limit of 75 topics because each of the models differs from the others enough to generate topics that may have been combined in single models with smaller topic numbers. Based on the keyword search and the algorithms for topic modeling and clustering, we generated our full corpus model ($N = 85,663$) along with seven disease-specific models. These models are not mutually exclusive at the word or document level; some words appear in multiple topics in any given model or across models, and some articles appear in multiple models. For example, it was common that a MERS article discussed and applied knowledge from the SARS pandemic.

2.4. External Validation: Randomized Human Annotation

To ensure that our modeling approach generated document clusters in a sense-making manner, we conducted an external validation. First, we randomly sampled 10,000 documents in our corpus and extracted their topics and clusters following the same modeling approach as described in the previous sections. Then, the clusters were assigned to one of the three categories (Bench Science, Treatment, Public Health) by a clinical subject matter expert. The definitions of the categories are listed below.

Bench Science literature investigates molecular mechanisms and markers of disease or therapies, for example, viral counts, LD50 (the median lethal dose), and protein expression.

Treatment literature covers clinical symptoms (fever, cough, hypoxia) or outcome-related measures, for example, length of stay, morbidity/mortality, intubation, or other respiratory support.

Public Health literature discusses measures with population as denominator.

Any unassigned clusters were labeled 'None.' Since each document can belong to multiple topics and multiple clusters, we listed only the top two clusters of each document based on the sum of the scores generated by the LDA topic models.

We annotated a portion of the sample documents ($N = 3,631$) on the same three categories to form a reference standard. We first prepared a training data set, which contained 250 documents from the sample. These training documents were categorized independently by two researchers and the discrepancies were discussed and resolved by a third researcher to generate the final categories. Then, this training set was handed to a group of research assistants, who reviewed the training documents and compared their categories with the final ones. Specifically, the research assistants reviewed 50 documents in each round of training. If a research assistant could achieve a 90% match, in other words, they matched the same category in 45 out of 50 documents, this person would be allowed to move on and annotate the nontraining documents in the sample. Otherwise, the

research assistant reviewed the documents and repeated the training process on the second 50 documents in the training set until passing the threshold. All human annotations were merged to form a reference standard.

Lastly, the reference standard was compared with the label of the first and second cluster generated by our modeling approach. Given a document, if the label of the first cluster matched the reference standard, it earned 1 point. If the label of the second cluster matched the reference standard, it earned 0.75 points. If none of the labels matched the reference standard, it earned 0 points. We calculated a percentage of agreement between the reference standard and the output of our modeling approach by summing all the points and dividing the sum by the number of documents in the sample.

The results of this external validation show that our modeling approach can achieve high agreement (71.03%), indicating the high validity of our approach. Since our modeling approach combines the strength of machine learning and expert knowledge, it is more efficient than conducting traditional human annotations to form reference standards on a large corpus. Moreover, our modeling approach can uncover hidden relationships among the documents and present the patterns in an easily interpretable manner through interactive network visualizations. Through this external validation, we demonstrated the efficacy of our modeling approach to support the results in the rest of the article.

2.5. Model of Models: Parallel LDA Replicates

For each of the seven virus models and the total CORD-19 corpus models, we run six parallel model replicates with different random seeds on each corpus, resulting in 48 models for analysis. We then visualize the resulting models' topics in a single network visualization per corpus. Rather than run a single model, which is an inherently variable process dependent on the initial Bayesian distribution, we propose running multiple models from different seeds and clustering the resultant topics using their word distributions to create more meaningfully stable clusters and to compare word usage across the parallel models.

These clusters, integrating topics from the six models into an aggregated model of models, serve both to confirm consistent topics across all models and to reveal underrepresented topics that may not have appeared in a single model representation. Each MoM run aggregates the results of six parallel replicates to stabilize our model outputs by identifying and boosting the signal of topics with overlapping vocabularies that recur consistently across the independent replicates. Correspondingly, the multiple simultaneous models indicate clearly topics with words that occur only rarely with faint representation in a single model, which will be less emphasized in the results. This distributed parallel approach aims to increase user confidence and the interpretability of our models by bringing the most frequently repeated topics to the top tier of the model results.

We also designed MoM's distributed parallel approach to LDA to prevent very common words or near-stop words from dominating topics by virtue of their nonspecific ubiquity. The dilution of topic meaningfulness by common filler words is a frequently invoked critique of classical LDA, which results from an overreliance on

the term-topic probability distributions as the primary result of the topic model (Blei, 2012). Anticipating this risk, we used both the term-topic and document-topic probability distributions to construct and evaluate our model of models.

Our analysis therefore evaluates topic meaning both in terms of the words and the documents populating them. Most studies employing LDA disproportionately favor the former while inadequately making use of the document-level information contained in the latter. Consequently, we compared both the term-topic and document-topic matrices in the construction of the two visualizations we presented to our human coders: the network analysis visualization of topic patterns and the information-retrieval interface displaying the underlying articles populating topic clusters.

Our visualization represents both word-topic distributions via location (node placement on the two-dimensional plane) as well as document-topic distributions via graph structure (edges linking nodes), allowing for a more layered presentation of the model results than merely showing one of these two learned distributions. We also allow for the hierarchical exploration of each cluster and topic's top words using a structured tree visualization. Combining the word- and document-level relationships result from the model of models allows us to check the strength of inter-topic relationships by evaluating the extent to which topics exhibit shared source documents in addition to shared words.

In our network, each node represents a topic in a planar vector space. The network locations are generated from a two-dimensional Principal Component Analysis (PCA) projection of each topic's l^2 normalized word distribution. We clustered these topics using the Elkan k-means algorithm implemented in the scikit-learn library, where k = the number of topics for the model (Pedregosa et al. 2011). The groups generated serve as 'topic clusters' at an aggregated model level, aiding interpretability.

Weighted links between topics in the graph were created using the number of documents in the corpus that are more than 1% likely to occur in both topics. While we represented these links on a sliding scale visually, so as not to render a meaningless, completely interconnected graph, we used the fully connected weighted graph to calculate shortest-path betweenness centrality using the NetworkX library to support our visual findings related to topic importance (Appendix A, <https://scholar.uc.edu/collections/mg74qn746>) (Hagberg et al., 2008). Measuring a topic's centrality to the network allows us to classify a topic's ability to serve as an information broker in the network, to the extent that information has a high likelihood of transmission through central broker topics that are densely interconnected within the network structure (Bakshy et al., 2012; Wilkerson et al., 2018; Vosoughi et al., 2015; Levine, 2009; Gruhl et al., 2004). In social network analysis, this centrality metric suggests influence (Everett & Borgatti, 1999). While our networks are formed around interactions of language, topics with high centrality retain that quality of influence (Sims & Bamman, 2020).

2.6. Subject Matter Expert Analysis and Agreement

We used a multidisciplinary, team-based approach to analyze and make sense of the resulting models. Nine domain specialists from infectious diseases, internal medicine, hospital medicine, biomedical informatics, biostatistics and epidemiology, library science, and linguistics evaluated the model outputs independently by tagging the themes associated with the words and documents contained within each topic and topic cluster.

Tagging proceeded in two steps. First, our DSC data scientists and digital librarians sorted the top words and underlying documents in each topic cluster. The DSC team and the clinicians then independently labeled the themes contained in them.

The results from each coder's reasoned judgment of the linguistic and document content underlying the models were compared, and topic labels were assigned based on full agreement between the coders. In cases of a single coder disagreement with the majority, we explained and discussed the reasoning process used to assign tags to the words and documents in a topic. Ironically, the COVID-19 pandemic prevented our team from assembling as a group, and so this facilitated the independent parallel tagging and agreement approach used to label our models.

We further analyzed the seven disease models by connecting these prevalent topic clusters with other topic clusters in the same model and then, in turn, across models for a comparative inter-virus perspective that took into account the thematic overlaps and sequential influences between the most notable viral outbreaks of the late 20th and early 21st centuries.

Our analysis of the COVID-19-specific model delves deeper into the question of how scholarly topics orient themselves in textual language by including a comparison of the to-date (as of July 14, 2020) scientific language concerned with the virus to that accumulated through March 30, 2020. These two snapshots of the data as it has evolved allows us to observe how virus/pandemic scholarly literature develops, further elucidating the ways in which knowledge builds upon itself and how the research community reorients its language based on that knowledge.

3. Seven Viral Outbreak Models (1980-2020)

3.1. On Reading a Topic Network

In a network of interpersonal relationships, connections between entities reflect social ties. When applied to a corpus of documents, our network visualizations of topics generated from a data set of scientific literature depict linguistic ties, namely, an overlap in documents. In the visualizations referenced here, each node in the network is a topic, assigned a color based on the cluster of topics to which it belongs. It follows that topic clusters have a hierarchical relationship to topics. Connections, or 'edges,' between these topic nodes are reflective of how many documents the connected topic nodes share, that is, the topics co-occur in N documents.

Given that this connection is based on shared documents, we can understand latent relationships that exist between topics at the level of shared research and, in turn, understand how and what information is synthesized and transmitted through those edges. Node location in the network, on the other hand, is based on how a topic's probability distribution of terms differs from other topics; topic nodes with common vocabulary will be closer to each other in the physical space of the network. When these topic nodes are closely related with a high degree of shared vocabulary, they coalesce to form distinct groups or neighborhoods.

These aspects of the network visualization make it possible for nontechnical team members to identify distinct clusters of topic nodes that share a similar vocabulary. The clusters visually focused our coders' attention on particular linguistic 'hotspots' that merited qualitative human analysis by topic tagging and reading of the underlying documents for verification of meaning. While an area densely populated by nodes forms the central focal point in a thematic neighborhood, a visibly dense accumulation of network edges signifies the degree to which nodes share common articles, and thus indicate document-level relationships between these neighborhoods and their topics. Our use of the network visualization's density of nodes and edges to highlight visually areas of linguistic and textual co-occurrence was a feature of our diverse team of clinicians and linguists who had little experience analyzing NLP models, and did not find value in poring over the matrices returned by our models. The network visualization served as a useful disciplinary and methodological common ground between our more technically minded team members and those who valued a graphical representation of the data for the purposes of identifying clusters and for information retrieval of relevant documents for review. Furthermore, the network visualization made the models more tractable for our student coders, who tagged and read a large volume of topics and documents for our external validation measures, described previously.

Between neighborhoods, additional content may exist that serves to connect related fields. From this comes the idea of bridging and brokerage. In particular, the connections outside of a strongly positioned neighborhood, or rather, the edges between neighborhoods, are generally weaker than those within the same topic neighborhood. The edges, however, are important in allowing information flow across the network as a whole and serve as structural bridges, ensuring the network's overall coherence (Granovetter, 1973).

When this visualization approach is applied to the virus models built in this study, we see distinct neighborhoods form repeatedly. These include Basic Science, Treatment, and Public Health neighborhoods tagged by our subject matter experts, creating a triangular network formation (e.g., Figure 2 and Appendices B, C, D). These network neighborhoods reflect the common translational science convention of research moving from "bench, to bedside, to population" or "laboratory, clinic, community" (National Institutes of Health, 2020). The connections between these neighborhoods provides insight into how these translational fields relate to one another in the response to viral outbreaks. Specifically, consideration of the bi-directional and unidirectional bridges between each neighborhood vertex of the triangular networks (Treatment and Public Health; Public Health and Bench Science; and Bench Science and Treatment) becomes critical to

understanding the composition of the corpus. The connections between these neighborhoods provide insight into how these translational fields relate to one another in the context of viral outbreaks. Without the bridging presence of certain topics, the network would break apart and neighborhoods would be isolated from each other in a way that prevents them from productively informing one another. In network analysis terms, the virus's topic networks rely on topics occurring within shared articles for connectivity and the transmission of information across the network (Granovetter, 1973). Drawing from Burt (1992, 2005) and Everett & Valente (2016), throughout our analysis we employ the definition of bridging as 'an edge property that measures the extent to which an edge forms a bridge,' while the related concept of brokerage is a 'node-level property' that defines 'control over bridging' throughout the entire network as opposed to a node's own neighborhood. Musial & Juszczyszyn (2009, p. 357) combines these two facets by defining network bridges as "the nodes without which the network will split into two or more subgroups."

Overall, there are two fundamental characteristics to consider in this type of network analysis: 1) The size of a topic cluster or topic depicting its weight in the overall network and 2) A topic's proximity to, and connections between, other topics. The primary focus of this article will be on the latter, aiming to better understand topic relationships and their modes of information synthesis.

Figure 2. Network visualization of the SARS Topic Model.

3.2. Weight & Connections: A Neighborhood Map View of Outbreak Models

The nodes included within the vertex defined as the Public Health neighborhood includes the terms ‘public’ and ‘health’ as well as vocabulary like ‘countries,’ ‘people,’ and ‘development.’ This neighborhood also tends to have a presence from either a data/modeling topic cluster, or several topic nodes with data analysis vocabulary. Though less immediately conspicuous, this neighborhood also tends to feature a smaller adjunct of transmission topic nodes. Public Health surfaced as a consistently present neighborhood for most models, except the MERS model, and is primarily formed by two or three dominant topic clusters in the virus models (e.g., Figure 3 and Appendices B, C). In three of the models (HIV, Zika, and Ebola) a Public Health topic cluster from this vertex neighborhood is also the first most prominent topic cluster in the network overall. This is possibly due to distinct public health methods, as modes of transmission differ. For the other four virus models (the three coronaviridae viruses, in addition to H1N1), the top prominent topic cluster belongs to the Treatment neighborhood. This is likewise true of the *total* corpus model ($N = 85,663$ documents) (Appendix D, <https://scholar.uc.edu/collections/mg74qn746>). Interestingly, the models where Public Health is strongest show Bench Science having a stronger presence in their models than Treatment, whereas models with Treatment as the strongest topic cluster show Public Health as next most prominent.

In examining the network’s triangular structure, we can observe that the spaces between each of the triangles’ vertices also exhibit relationships between Public Health, Treatment, and Bench Science. Between the Public Health and Treatment neighborhoods, we observe denser edge connections, suggesting shared research between topics in those areas. On the other hand, the space between Public Health and Bench Science shows sparser connections while simultaneously containing more topic nodes, depicting a more diverse shared vocabulary between these neighborhoods despite less shared research published. Lastly, the bridging space between Treatment and Bench Science appears relatively empty compared to the other two bridging spaces, with an observable lack of both connections and nodes (e.g., Figure 2). This inner-network space will be discussed in greater detail in Section 4.2. Overall, this ‘bird’s eye view’ of node and edge patterns between neighborhoods (or vertices) is especially strong across the respiratory virus (SARS, MERS, and COVID-19) and H1N1 networks in particular.

In analyzing the virus-specific models from outbreaks occurring before COVID-19, we still observe COVID-19 documents and language in our networks (Figure 1). This shows, in some part, the motivation of our data set. It also makes evident the effect of earlier research on the current pandemic. For example, the sub-data sets for SARS and MERS (the other coronaviruses) decrease the most when we exclude COVID-19 documents, which meets expectations based on the viral similarity between these outbreaks. Regardless of virus, the presence of COVID-19 remains the same across the models, with a position in the Treatment neighborhood. When we do not exclude COVID-19 documents from the other coronavirus models, we see a much more robust Treatment neighborhood across coronavirus models. For both the SARS and MERS models, this accounts for much of the decrease in documents we see when we exclude COVID-19 documents from these

data sets. All noncoronavirus models depict Treatment as the least prominent neighborhood in the network visualization, whether or not COVID-19 articles are excluded, but again, COVID-19 is observable primarily within the Treatment neighborhood.

3.3. Coronavirus Models (SARS, MERS, and COVID-19)

Human coronaviruses typically cause mild and transient respiratory or gastrointestinal illnesses. Until recently, they were considered only minor human pathogens. In 2003, however, a novel coronavirus (SARS-CoV) was identified as the cause of Severe Acute Respiratory Syndrome (SARS), which is characterized by pneumonia with progressive hypoxia and respiratory failure. There were approximately 8,000 cases reported in 32 countries, with a mortality rate of approximately 11%.

When we analyze the SARS network model (generated from a data set of 30,382 documents), we found that it prominently features a Public Health topic cluster in its model, Cluster-34 (*health-information-development-people-research*) along with a Data/Modeling topic cluster enmeshed in this neighborhood, Cluster-30 (*model infected individuals cases data*). Together, these two topic clusters form roughly half of this vertex space. Despite a prominent Public Health neighborhood however, the SARS model still cedes the top spot to Cluster-2 (*patients-covid-19-severe-patient-risk*) containing language aligned with Treatment (Figure 3). This is the largest topic cluster in both subtopics and document count, twice as large as the second cluster by both metrics. Unlike the other respiratory models, this lone cluster dominates an entire vertex of the network, although Cluster-32 (*patients-patient-pandemic-staff-covid-19*) is close. Nonetheless, what might be labeled the Treatment neighborhood here also exhibits COVID-19 vocabulary. Despite the particular inclusion of COVID-19 in the SARS Treatment neighborhood, this neighborhood in the SARS network will prove to behave in the same way as the Treatment neighborhoods from other virus models; containing Treatment vocabulary, forming multiple, strong connections to the Public Health neighborhood, and featuring a relatively uninhabited bridging space between itself and the Bench Science neighborhood.

The strong influence of COVID-19 in the Treatment neighborhood of the SARS network clearly reflects the foundation of our data set (CORD-19), which is highly focused on COVID-19 literature. In this example, excluding COVID-19 from the SARS model cuts 55% of the articles from the data set. It is interesting to note here that this COVID-19 influence is less intense in the MERS network than in the SARS network, and weaker still in the models for noncoronaviridae (Figure 1), suggesting a more limited impact of MERS and noncoronavirus literature on COVID-19 research. The most significant area in which SARS and COVID-19 literature overlap can be observed in the Treatment neighborhood, which is dramatically weakened without COVID-19 documents, shifting the SARS model's primary focus to Public Health. This may reflect two phenomena: 1) SARS, unlike COVID-19, never reached the pandemic stage and was eliminated before the need for widespread therapeutic efforts and 2) the established understanding emerging from the SARS literature formed a foundation for researchers looking for COVID-19 therapies, thus making the overlap in Treatment-related topic clusters when the SARS models did not explicitly exclude COVID-19 documents. In

examining both versions of SARS networks (with and without COVID-19 articles), we still observe a similar lack of activity between the Treatment and Bench Science neighborhoods. Furthermore, the topic nodes for each model version that do surface in that space and serve as bridges between the neighborhoods are notably similar, containing vocabulary such as “increased, expression, production, patients, mice, and levels” (Figures 2 and 3). The SARS model, excluding COVID-19 documents (Figure 3), also more clearly defines an offshoot from the Public Health neighborhood made up entirely of vocabulary indicative of zoonotic/transmission studies, Cluster-6 (human-humans-animals-disease-transmission). This is unsurprising considering one of the main public health discussions following the emergence of SARS revolved around natural reservoirs of coronaviridae and mitigation of transmission events to humans.

Figure 3. Network visualization of the SARS Topic Model excluding COVID-19 documents.

Almost a decade following the SARS outbreak, another novel coronavirus (MERS-CoV) was identified as the cause of the Middle East Respiratory Syndrome (MERS), which resulted in over 1,000 cases, concentrated primarily in the Middle East and South Korea. In comparison to SARS, the transmission of MERS was less pronounced, with fewer countries and fewer individuals affected. It is thus interesting, but perhaps not surprising, that the MERS model (generated from a data set of 8,855 documents) depicts Public Health as much less prevalent. Despite the difference of weight in the networks' three neighborhoods, the MERS model still shares many features in common with the SARS model. Like SARS, the MERS network depicts the strongest connection between Public Health and Treatment neighborhoods in the network (Figure 4).

Both the MERS and COVID-19 networks divide their Treatment neighborhood vertices between two topic clusters, with one topic cluster's vocabulary leaning closer to Bench Science and the other's toward Public Health. However, the MERS model all but eliminates this Treatment neighborhood once COVID-19 documents are removed, similar to what we observe in the SARS model and likely for the same reasons (Figure 5). Thus, only when COVID-19 documents are explicitly excluded do we see Public Health topics strengthen in the MERS model. Despite something of a redistribution of weight, this Public Health neighborhood of the network excluding COVID-19 remains more diffuse, less clearly defined than what we observe in other models. Notably, the zoonotic/transmission topics we saw as an offshoot group in the SARS model exist more with this space, mirroring the significant discussion about zoonotic transmission throughout the MERS outbreak. In contrast to MERS and SARS, both of which are responsible for outbreaks that have since been declared as resolved, the resulting pandemic caused by a third novel coronavirus (SARS-Cov-2), which causes COVID-19, is still on-going. Therefore, the COVID-19 models (Figures 6 and 7), are intrinsically different since this pandemic continues to evolve. This suggests that, while the similarities of this model are even more notable, given that they surface early in the outbreak timeline, the distinctions may simply reflect the snapshot-in-time nature of its underlying data set.

Figure 4. Network visualization of the MERS Topic Model.

Figure 5. Network visualization of the MERS Topic Model excluding COVID-19 documents.

With this in mind for the COVID-19 research evolution, we analyzed two points on this outbreak's timeline; one model depicting the earlier response through March of 2020 (Figure 6), then another, fuller model through July of 2020 (Figure 7). Both models still align with the three-neighborhood network structure, featuring vertices from Bench Science, Treatment, and Public Health. While Treatment topic clusters form the strongest neighborhood in both models from the COVID-19 timeline, they outperform the other neighborhoods in the earlier research response to a stronger degree consistent with early efforts to identify patient-level risk factors or effective therapeutic options for this potentially fatal disease. These topics are present in Cluster-11 (patients study severe days patient) and Cluster-4 (patients patient risk pandemic severe).

learning || machines

Figure 6. Network visualization of the COVID-19 Topic Model as of 03/2020.

Public Health, at this earlier point as well, has a strong presence, though to a lesser degree than Treatment (see Cluster-1 [data model infected epidemic china] in Figure 6). Interestingly, we can observe Data/Modeling topics taking command of this neighborhood earlier on, rather than being featured as a secondary subcluster in the neighborhood like we see in other, more complete models. In the context of dominant Treatment topic clusters, it is noteworthy that the Bench Science neighborhood is noticeably less prominent in the COVID-19 models, as well as less focused than what we will show to be typical in the other virus-specific models. Its distinct position in relation to the rest of the network is consistent with other models, but the Bench Science presence suggests an initial scattershot effect where efforts to understand this novel virus were highly variable in both focus and approach. This stands in direct contrast when compared to a more established disease model, like HIV, where longstanding research has contributed to robust understanding of viral mechanisms and allowed for selection of the most promising targets for therapy. The contrast might be best understood as the evolution of basic science research approaches from emerging to established viral pathogens, which begins as an unselective scattershot approach but focuses dramatically over time as the most promising strategies are adopted by the research community.

Figure 7. Network visualization of the COVID-19 Topic Model as of 07/2020.

It should be kept in mind that network connections across neighborhoods are weaker overall in the earlier COVID-19 model, specifically connections to Bench Science, which leads us to understand that Bench Science research stayed more siloed in the early response to this pandemic.

3.4. Other Noncoronavirus Epidemics

Within the CORD-19 data set, we investigated several other noncoronavirus viral outbreaks. These other viruses included, chronologically: human immunodeficiency virus (HIV), Zika, influenza A H1N1 (H1N1), and Ebola (see Figure 1). This additional analysis adds a historical context to viral outbreak research, as well as external validation using epidemics and pandemics unrelated to coronavirus virology and clinical characteristics. It is important to note, however, that because the articles included here were selected based on their relationship to COVID-19 themes, their inclusion in this data set is not reflective of complete bodies of their research literature. In other words, the models of these viruses are generated from somewhat limited documents, and not reflective of the complete realm of scientific literature about them. Despite this limitation, identifying similarities and differences between coronavirus and noncoronavirus models may help validate our conclusions and provide preliminary data to support future applications of this approach to more complete data sets.

Human Immunodeficiency Virus (HIV): The model of the HIV data set ($N = 12,024$) is somewhat unique among our collection (Figure 8). Though we still observe our three network neighborhoods, the HIV network

has a weaker Treatment neighborhood than the coronavirus models, and the strongest Bench Science presence among all of the virus models (both coronavirus and noncoronavirus), primarily formed around four clusters that are all prominent in the model:

- Cluster-15 (*cells-results-infected-incubated-expression*)
- Cluster-7 (*expression-proteins-viruses-translation-mrna*)
- Cluster-3 (*binding-proteins-protein -interaction-peptides*)
- Cluster-14 (*expression-mice-cells-human-production*)

Public Health follows in weight, and here we observe another difference, that its Public Health neighborhood consists of topic nodes from multiple clusters, in a more dispersed fashion, rather than a few dominant topic clusters. This Public Health neighborhood also has something like a transmission attachment. We have seen this in other models as well (SARS and H1N1), suggesting the influence of their epidemic status by their proximity to larger scale public health topics. For the HIV model, however, we see the term ‘outbreak’ surface only to the topic-cluster level—not ‘pandemic’ or ‘epidemic.’ This may reflect the scientific community’s perception that tackling HIV at the public health level has evolved past epi- or pandemic strategies and, at the local level, consists of understanding and mitigating outbreaks.

We hypothesize that these differences in the HIV model are based on several factors, including the longevity of the HIV epidemic, leading to more mature understanding of the virus, its underlying mechanisms, and established therapeutic interventions. Considering these differences, it is not surprising to see that the HIV network’s Bench Science neighborhood outweighs the other novel infectious disease networks. Since HIV’s emergence in the 1980s, considerable effort and resources have been expended to understand the pathogen and effective public health and pharmacologic interventions (Forsythe et al., 2009). This stands in direct contrast to MERS, SARS, and COVID-19, which have emerged more recently and, until now, have not received widespread attention from the biomedical community. Another of the many distinguishing features of HIV is that the primary route of morbidity and mortality is secondary, meaning that the virus inhibits the immune system’s ability to fight off otherwise routine pathogens and HIV-infected patients die of other infections (Croxford et al., 2017). This is one reason other infectious diseases appear in the HIV Infection model’s topic clusters but are absent from other viral outbreak models. This might also be why we observe, similar to the other viruses, a large void area between the Treatment and Bench Science neighborhoods in the HIV model’s network yet miss a clearer presence of broker or filtering activity happening between these neighborhoods that we see in other networks.

Figure 8. Network visualization of the HIV Topic Model.

Zika Virus: The Zika model network (generated from a data set of 1,096 documents) features its Public Health neighborhood as most prominent. Similar to the MERS model, we see topics related to transmission playing a strong, enmeshed role in this neighborhood, unlike the adjunct status of these topics in the SARS and H1N1 models (Figure 9). This may be related to the unique transmission modes of Zika virus compared to other viruses compared in this analysis. Zika virus is an arbovirus, meaning that it is transmitted by an arthropod. Human-to-human transmission is also possible, but most commonly through sexual intercourse or from a pregnant woman to her fetus. While Zika virus had been identified decades earlier, the larger outbreak in 2015–2016 was associated with a very specific complication: microcephaly in neonates born to mothers infected with Zika during gestation. Due to these multiple routes of transmission, and concerns for specific complications of infection, unique Public Health interventions to prevent further spread of disease were and are required. This directly influenced the outbreak response, as well as its impact in informing interventions.

Figure 9. Network visualization of the Zika Topic Model.

Influenza A Virus Subtype H1N1: In the grouping of coronavirus and noncoronavirus models, H1N1 holds a particular place. While it is not a coronavirus, it is still a respiratory pathogen, and that similarity is reflected in its model, which shares some patterns with each group (Figure 10). In the H1N1 model ($N = 12,720$), we observe a network more closely resembling the coronavirus models, with a densely populated space between Public Health and Bench Science that is inhabited by a multitude of smaller topic nodes. Additionally, connections between Treatment and Public Health neighborhoods are stronger and, unlike the other noncoronavirus models, the Treatment neighborhood is pronounced. In fact, the two most prominent topic clusters in this model contain a Treatment vocabulary. However, as we observed in the Zika network with its Malaria/Treatment cluster, there exists a second Treatment component in the H1N1 network between Public Health and the stronger vertex of Treatment. The strength of Treatment topics in this model, then, is in part due to this topic cluster combining Treatment vocabulary with a specific focus on children and asthma, Cluster-12 (children patients asthma rsv age) highlighting a specific research focus that targets the confluence of two risk factors for severe infection secondary to respiratory viruses: young age and asthma. Based on precedents and experience with Respiratory Syncytial Virus and Influenza, it is unsurprising to see this topic emerge as a focal point in literature discussing the novel H1N1 strain of influenza.

Figure 10. Network visualization of the H1N1 Topic Model.

Figure 11. Network visualization of the Ebola Topic Model.

Ebola Virus: Public Health is the most prominent topic cluster in the Ebola model (generated from a data set of 4,089 documents) (Figure 11), though it also features a reasonably prominent Bench Science neighborhood, like the other noncoronaviruses, and unlike the coronavirus models. Topics within this Bench Science neighborhood connect most strongly within their own neighborhood, though gradually, as the threshold of shared documents is lowered. Public Health clusters begin connecting through zoonotic transmission topics in Cluster-9 (species humans bats viruses virus) to Bench Science and from Public Health to Treatment through topics in Cluster-3 (patients patient transmission risk cases). The reliance on these transmission clusters to connect different neighborhoods reflects this virus's notable high risk of transmission through contact with contaminated bodily fluids. Like the HIV and Zika networks, there are more network edges (representing shared research articles) existing between topics from Public Health and Bench Science, rather than the stronger tie between Public Health and Treatment areas we observe in coronaviruses and H1N1.

4. Outbreak Trends Among Viruses: Comparing Recent Historical Virus Models From Cord-19 Data to Covid-19

4.1. Reading Outbreak Trends Across Topic Networks

Based on our observation in the individual models, each of the models are defined by three major topic neighborhoods: Basic Science, Treatment, and Public Health. These neighborhoods are characterized not only by the strength of the ties between their topics, but also by the ties connecting them to other neighborhoods. The more subtle connections discussed in individual virus networks gain value and meaning when we see the same patterns recurring across multiple different virus models. This repetition becomes especially meaningful for understanding the temporal evolution and structure of scientific research and its dissemination in a specific domain. There are structural and thematic unities repeated across the models for the seven viral outbreaks, and these similarities highlight the ways in which the CORD-19 data set—a data set reflecting research primarily concerned with an ongoing outbreak—deviates from the patterns established by the six notable outbreaks that preceded it. We draw distinctions specifically between COVID-19 and the other coronaviruses in the data set because SARS and MERS are the most directly comparable viruses, as previously explained. We posit that the pervasiveness of these patterns is what makes the established, historical virus networks so cohesive, and allow for the mediation of information among these dissimilar neighborhoods.

4.2. Bridging Bench Science and Treatment

The space between Bench Science and Treatment neighborhoods is all but empty, with a notable absence of topic nodes depicting language applied equally to both neighborhoods. Investigating this further, we examine the edges formed between topic nodes from the two neighborhoods, depicting shared documents and shared research. This relationship too, though, is similarly faint. By lowering the threshold of shared documents required to form edges between topic nodes, however, we can see a first connection between the two neighborhoods finally surface, appearing to concentrate that contact through a first, broker connection. We

suggest that this early pinpoint, like a concentrated light source, is the first glimmer of the scientific community beginning to centralize around a common theme that may bridge the gap between basic and clinical science, bench and treatment. This reflects the filtering process imposed by the scientific and medical communities themselves, between bench science and clinical application. For a discovery in bench science to advance into the treatment neighborhood, specific criteria must first be met, beginning with preclinical studies, and progressing toward controlled clinical trials. This progression is time-consuming, expensive, and associated with significant risk. Therefore, it is usually not until sufficient evidence accumulates that the field begins to accept a topic as having potential clinical merit and worthy of more concentrated focus.

This structural pattern is present across models, whether or not their data sets include COVID-19–related documents, demonstrating that this theme of emerging scientific concentration bridging the basic and clinical science realms is universal in outbreak research. The specific nuances of each point of concentration, however, are difficult to discern based on topic nodes alone. Therefore, we interrogated the models more closely to read the underlying documents from which these topics were formed. In doing so, we found that topic and topic-cluster language did provide an accurate overarching theme for these concentrating connections, but the specific context of which was more readily defined through document review. For example, in the SARS model, excluding COVID-19 documents (Figure 3), the point of concentration focuses on terminology surrounding *in vitro* and *in vivo* models that explore the immune response to respiratory infections. More specifically, these are mouse models of respiratory disease and the effect of pulmonary infection on the cellular level, namely, the effect of various components of the inflammatory pathway, represented by the topic (mice expression cells increased production). When evaluated, the documents themselves that made up this node, the top articles are all related to innate and adaptive immune responses. Though we can see a similarly concentrating node in this space of the SARS network that does *not* exclude COVID-19 documents (Figure 2), a closer examination of the articles most related to this topic depicts an undeniable focus on COVID-19. In this SARS model that includes COVID-19 documents, this concentrated node appears to deal primarily with the characterization of risk factors associated with severity of disease in COVID-19. The articles depict a mix of predictive-oriented work alongside retrospective review with the goal being the same: How can we help clinicians risk stratify who is going to get sick and require intensive care versus who is not? We previously pointed out that most of the influence of SARS research on the current pandemic surfaces in Treatment topics. Here, we can see that this influence extends into this connection between these two neighborhoods as well, through the network’s node that illustrates the filtering process prior to Bench Science’s transition to Treatment.

In the MERS, excluding COVID-19, documents (Figure 5), the network’s concentration node has language focused on animal models of disease (mice, days, day, cats, increased); the top documents contributing to this node all clearly reflect an effort to better understand the pathology of and potential therapeutic options for MERS. Such preclinical trials are essential for the drug development pipeline, so their appearance as a concentrated focus node in the space between bench science and treatment is not surprising. When we look at

the MERS, *including* COVID-19, we see similarities to the SARS, including COVID-19, where the top documents in the concentrating node revolve around the role of immune response to respiratory disease.

Despite a much smaller sample size, we also saw this translational research field concentration process occur even in the model depicting the earlier response to the current COVID-19 pandemic (2,192 documents from a 3-month period, January through March) (Figure 6). However, this pinpoint topic node is located in a less stable, scattershot Bench Science neighborhood and included only pathogen-identifying language (sars-cov-2 human virus patients sars-cov). It is striking to recognize that, at that point in the pandemic, this was the most pinpointed language that the research community shared in the space between basic science and treatment. Upon interrogating the documents within this node, we identified a diverse number of topics, including a variety of potential preexisting treatment options and general discussions of COVID-19 epidemiology. Again, this reaffirmed the fairly broad nature of research at the time, as investigators explored in all directions and were still making an effort to understand the basics of the disease. Not surprisingly, when we expanded to the more mature COVID-19 model (Figure 7) we see more specialized language begin to surface in this role, focusing on terminology related to cytokine-mediated immune responses (patients il-6 inflammation increased cytokine), which appear to be critical components of disease severity and patient outcomes (Giamarellos-Bourboulis et al., 2020). This was apparent in the related documents, which reference cytokine storms, immune dysregulation, and immune modulation as a potential therapeutic option.

In the noncoronavirus models, we see similar phenomena related to the bridging space between bench science and treatment. Many nodes located in this space contain virus machinery and mechanism-related language, discuss mechanisms of pathogenicity, and float targets for vaccine or therapeutic intervention (see: Cluster 1, in Appendix C-8, <https://scholar.uc.edu/collections/mg74qn746>). Similar to the coronavirus models, by locating the specific topic node that carries the weight of connecting Bench Science to Treatment, we can better characterize the scientific community's perceived bridge from viral pathogenesis to legitimate strategies for safe and effective treatment. Furthermore, by analyzing this space across viral outbreaks, it allows us to better understand the central themes of distinct research responses.

4.3. The Evolution of COVID-19 Research

While viral outbreak research is continuing, the virus models reflected in our study (excluding the COVID-19 model) are based on research that we can consider as relatively stable, with a reasonable plateau given the historic nature of their respective outbreaks. As a pandemic still in progress, the data set of COVID-19 research provides a unique opportunity to visualize how Bench Science, Treatment, and Public Health research interrelate as the scientific community's understanding of a virus and its associated pandemic evolve over time.

While the Bench Science neighborhood of the COVID-19 model remains scattershot, even in the later model for this pandemic, more specialized vocabulary becomes more apparent over time. There is a movement away from the general terms occupying that space in the earlier model, such as 'sars-cov-2,' 'human,' and 'virus.'

Furthermore, there is an evolution within the specialized language itself, suggesting a focus on particular themes as a pandemic progresses. For example, two more specialized and much smaller Bench Science topic clusters that exist in the earlier COVID-19 model, Cluster-14 (epitopes patients peptides outbreak day) and Cluster-12 (ace2 binding medical waste structure rbd), are not necessarily represented in the same way as similar topic clusters in the later stage model. These earlier clusters may reflect initial broad discussions revolving around proposed mechanisms of disease and potential targets of therapies that dominated the conversation early in the pandemic phase, which are modified as more information is gained about pathogenesis and effective treatment approaches.

This evolution of research is particularly noteworthy in the context of one particular treatment approach that was explored for COVID-19. In the first COVID-19 model (Figure 6) ($N = 2,192$), which represents research published from January through March of 2020, Cluster-0 (patients sever treatment chloroquine studies) includes 10 subtopics, one of which is the full term ‘hydroxychloroquine.’ We see this cluster appear in the network just outside the vertex of the Treatment neighborhood, trending slightly toward Bench Science. In the second quarter of 2020, however, we see this hydroxychloroquine term move away from the Treatment neighborhood and closer to both Bench Science and the centroid of our triangular network. It appears in a less prominent topic cluster, reflecting the question of its efficacy, Cluster-0 (treatment drugs efficacy trial patients) (Figure 12). Finally, when we evaluate the complete COVID-19 model, including 31,818 documents from January through July 14, 2020 (Figure 7), we see ‘Hydroxychloroquine’ as a top term in a prominent topic cluster (Cluster-20 [treatment drugs patients efficacy hydroxychloroquine]), but one that is clearly self-isolated when we examine the edges of the topics. This evolution of hydroxychloroquine through the network over time reflects the tumultuous discourse regarding its evaluation as a therapy for COVID-19, culminating in an early conclusion of clinical trials due to an interim analysis that indicated that there was a lack of efficacy. Early in the pandemic, this existing drug was thought to be a potential therapy to reduce virus-related mortality. What followed amounts to a divergent discourse, where the early promise of this drug was politicized while scientific discourse raised serious doubts about its efficacy. We suspect the decline in prominence, along with the isolation of this cluster, reflects the sustained and focused discourse about hydroxychloroquine’s merit as a therapy despite the broader scientific community moving on to other, more promising approaches to treatment.

Not only do we observe a distinct evolution of research within the areas of Basic Science and Treatment, but also within the context of Public Health and Treatment. For example, in the earlier COVID-19 model, ‘pandemic’ appears as a top term in one Treatment topic cluster (patients patient risk pandemic severe), while another includes ‘epidemic,’ and two more include ‘outbreak.’ After a further three months of research output, however, the number of topic clusters with ‘pandemic’ as a top term has increased to four and we see that term begin to surface in more diverse topics across both Treatment and Public Health neighborhoods. Additionally, we see the inclusion of ‘pandemic’ alongside language related to student impact, mental health, and economic terms. We postulate that the heterogeneity of language early on reflects the community’s attempt to characterize the threat of COVID-19 as it emerged and prior to the formal declaration of a pandemic. This

evolves over time, with the emphasis not only shifting to terminology associated with a pandemic, but also to that which has a broad impact on public health and well-being. Interestingly, concurrently with the more global ‘pandemic’ language that we observe over time, we also see a decline in emphasis in the term ‘China,’ which appears as a top term in 5 of the 15 topic-clusters (33%) for the earlier-stage COVID model and only 3 of the 40 (7.5%) when we look at the model that includes later research. Again, this reflects the evolution of the scientific community’s geographic focus as the virus emerged from local context and onto the world stage.

Figure 12. Network visualization of the COVID-19 Topic Model from 04-07/2020.

Finally, it is interesting to note that in the early stages of the COVID-19 research response, we see “social media” appear as a top term in 2 of the 15 topic clusters. This drops off the topic clusters top terms in more mature models, but depicts a particularly contemporary issue for researchers when responding to a general public’s initial reaction to a scientific challenge and health crisis. It remains to be seen how social media has an impact on future scientific discourse and public health response.

4.4. Centrality: Topics and Terms

The visualization of the networks’ structure allows us to not only identify language neighborhoods that reflect three areas of research (Bench Science, Treatment, and Public Health) but also allows us to observe that those areas have distinct interrelationships. In other words, the research connecting topics of Public Health and Bench Science operates very differently than research connecting topics of Bench Science and Treatment in terms of both shared vocabulary and shared documents. Given the triangular structure of these models, we

examined vertices / neighborhoods in pairs through their connections. However, there are also terms and topics that unite the disparate neighborhoods. Therefore, to understand the mediating role of topic nodes that are most connected to the entirety of the network, we examined the most central topics based on their edge connectivity or betweenness centrality. When we turn our attention to betweenness centrality, we can observe one perhaps less conspicuous way that the COVID-19 motivation of our data set affects modeling. In networks that reflect a virus data set that excludes COVID-19 documents, we see higher centrality scores among the topmost central topic nodes than those that include COVID-19 referencing documentation (Appendix A, <https://scholar.uc.edu/collections/mg74qn746>). When we include COVID-19 documents, these most central topics are more likely to include COVID-19 vocabulary. This suggests that corpora centered on one specific virus have more personalized endemic vocabularies that are more likely to co-occur across the triangular field. This is confirmed by the fact that the model reflecting a randomly selected 10,000 documents also has lower centrality scores for its topmost central topic nodes, and that the topic nodes in the model of the total corpus are lower still. The more viruses included in the corpus, the more diluted the central topics become.

We hypothesized that topic nodes containing terms most endemic to our data set would have higher centrality measures. This was confirmed through qualitative analysis of the topics and terms included in the 10 highest centrality–scoring nodes in all models. These most central nodes contained pandemic- and virus-specific language such as ‘virus,’ ‘infect,’ and ‘case,’ as well as general research language like ‘study’ and ‘data.’ Interestingly, however, terms and topics associated with Treatment, such as ‘patient,’ ‘treatment,’ and ‘human,’ appeared in most central nodes more ubiquitously than did terms or topics associated with Bench Science or Public Health. Furthermore, Treatment research is often referenced by Public Health and Bench Science research as well, serving as something of a vocabulary bridge between the other two neighborhoods. Overall, this is perhaps unsurprising, considering the ultimate goal of the scientific community’s research efforts is to translate insight into strategies used to treat patients suffering from any given ailment.

5. Conclusions

5.1. How Network Bridges Differ Between Neighborhoods

The network modeling that we completed in this study by combining natural language processing, qualitative human coding of topics and documents, and mixed-methodology approaches to evaluate the COVID-19 research literature has allowed us to clearly identify three key components of scientific discourse surrounding viral epidemics: Bench Science, Treatment, and Public Health. The findings reported here suggest a linear relationship between two of these neighborhoods, with information moving from Bench Science to Treatment in a chronological sense. Indeed, this linear relationship is consistent with the translational research paradigm, in which the ultimate goal is to move scientific discovery from the ‘bench’ to the ‘bedside’ and, ultimately, the ‘population.’ The connections between other neighborhoods are more nuanced, however, and do not necessarily reflect such a purposefully linear relationship. Stronger and more frequent connections between nodes in the Public Health and Treatment neighborhoods inversely reflect a more insular Bench Science

research response. Furthermore, the ray bridging the Public Health and Bench Science vertex neighborhoods contains a multitude of tiny topic nodes forming something that looks like a pebble path, reflecting a more specialized vocabulary and topics that are not the primary focus of research. We suspect that unfocused areas such as this may also be reflected in the public domain as well. For example, during the current COVID-19 pandemic, Public Health messaging has been incoherent or inconsistent, with limited connection to Bench Science research. This has been most apparent in terms of mask wearing, where advice has either been contradictory or inconsistent (Leung, 2020). We speculate that observing such ‘pebble path’ network connections in future models may serve as an indicator for similar inconsistency in other outbreak responses.

5.2. Scale and Urgency

In addition to the durable connections of Bench Science, Treatment, and Public Health research in the context of viral outbreak response, our findings also highlight the urgency with which this research has been pursued in the context of a pandemic. By evaluating the evolution of our COVID-19 model, we demonstrated how the scientific literature shifted from language of an outbreak, to an epidemic, and, finally, to that of a pandemic. Throughout this transformation, the sheer volume of literature that was generated reflected a cacophony of voices hoping to share perspective, advance understanding, and contribute to a solution to the pandemic. We posit that the scale and scope of the scientific community’s response is directly related to the global sense of urgency that developed as the devastating impact of the virus became apparent. For the COVID-19 response, we can see this clearly reflected in topic clusters that depict the pandemic’s reach into mental health, education systems, family management, and the economy:

- Cluster-36 (anxiety depression pandemic stress mental health)
- Cluster-7 (pandemic people crisis impact students)
- Cluster-15 (children parents adults families child)
- Cluster-26 (countries production consumption china pandemic)

Such language, particularly related to mental health, is absent from other viral outbreak/epidemic models, with the exception of ‘depression’ appearing within a single, smaller topic once each in the SARS and MERS models in the context of survivors/PTSD. This may reflect the scientific community’s growing acknowledgement of mental health’s legitimate role in contributing to human well-being, especially in the context of a pandemic.

With the lack of clearly defined treatment guidelines for COVID-19, clinicians are instructed to rely upon conclusions from or use therapies only within the context of clinical trials of novel interventions (Bhimraj et al., 2020). In response to the global pandemic, clinical trial research related to COVID-19 treatment has emerged at an unprecedented rate (Thorlund et al., 2020). In the context of a lack of a Treatment network bridge in the COVID-19 model, which reflects an ongoing outbreak in the information era, it is possible that early or preliminary signals in biomedical research are amplified more quickly and dramatically than in the

past. At present, where the internet and digital media are frequently used to communicate medical information, whether it be accurate or not, there may be a lower threshold used for broadcasting conclusions based on small or poorly designed studies conducted in emergent clinical environments. In the absence of these contemporary information dynamics (where a research finding can be communicated and broadcast regardless of its rigor, as may have been the case in historical viral outbreaks) it is reasonable to expect a more cohesive, organized voice to emerge from a broader, more conclusive body of evidence. For this reason, it is possible that, even as the COVID-19 literature matures and our understanding of both disease mechanism and effective treatment modalities advances, there may never be a well-defined Treatment topic cluster rising above the chaos to serve as the information bridge between Public Health and Basic Science as was observed in prior viral outbreak models. Continued analysis of the COVID-19 literature is essential, however, to test this hypothesis.

Just as the evolution of pandemic language, and the inclusion of unique terms related to broader states of wellness, provides insight related to the scale and urgency of the pandemic, the overall analysis of the COVID-19 network as it matured over time proves to be enlightening. With specialized language and topic clusters being more prominent at different time points, we can identify areas of emerging promise and, in the case of hydroxychloroquine, fading expectations. This type of literature synthesis may help the research community better understand the landscape of basic science approaches and, in some cases, the most promising routes to therapeutic breakthroughs. Not only does this provide context through which we can retrospectively evaluate a research response, but it may also help inform funding bodies on where and how to focus resource expenditure to contribute to meaningful breakthroughs. Furthermore, these approaches may aid investigators in identifying underleveraged approaches or opportunities for exploiting a gap or white space in the landscape of approaches for a given disease, rather than committing focus on an already oversaturated field.

5.3. The Value of Convergence Research for the COVID-19 Crisis

By filling the gaps in a research landscape, scientists may help identify areas that are central to the domain of the highest convergence and, thus, serve as a bridge between multiple disciplines (National Science Foundation, 2016). We propose in the current COVID-19 crisis that investigators capable of synthesizing cross-disciplinary information are perhaps the best candidates for (1) strengthening the bridge space among Bench Science, Treatment, and Public Health and (2) establishing research around the most central / convergent topic clusters throughout scholarly publications. It is, after all, these kinds of research teams whose publications populate not only bridge spaces, but also the other convergent topic clusters found to be most central in the other coronaviruses (Figure 2 and Appendix B, <https://scholar.uc.edu/collections/mg74qn746>). Teams of clinicians and scientists who value research convergence in this way can themselves act as brokers among the components of the Bench-Bedside-Population pipeline because they, and the teams they work with, have access to a wide array of subject-matter expertise. Research convergence in scientific academic literature will both build stronger bridges in virus research networks and provide more harmonized and coherent information to link normally disparate subdisciplines concerned with the COVID-19 pandemic. Convergence research,

therefore, because it is positioned at critical locations in the research networks and functions centrally in connecting the articles in the scientific literature, is imperative to epidemic outbreak research. Setting a research direction more intentionally toward the goal of convergence research could serve to accelerate the pace of discovery toward effective COVID-19 Treatment, with the ultimate goal of resolving this contemporary global crisis.

Our own research team implemented such a convergent multidisciplinary approach in miniature in the course of this study. Our project employed the novel collaboration of investigators from infectious diseases, hospital medicine, internal medicine, biomedical informatics, library and information science, computer science, and linguistics. Our goal was to combine methods and knowledge from these disparate fields to perform a more kaleidoscopic analysis to embody the convergent research practices we propose at a greater scale. Intense discussion and debate between these collaborators at every point in the research process served to elevate the discourse and provided critical links between vocabulary and texts, concepts and trends, and information synthesis.

6. Future Directions

The COVID-19 pandemic is ongoing, and the literature surrounding this global crisis is rapidly developing. Undoubtedly, the network model will continue to mature and evolve in ways that are both predictable and unexpected. Ongoing review of the COVID-19 data set's evolution will provide further insight into how scientific discovery unfolds over the course of an epidemic. Furthermore, by comparing viral disease network models throughout the course of an outbreak, rather than as a series of snapshots in time as described in this report, we may also begin to define metrics for success and failure in terms of a scientific community's research response to a crisis. By developing such metrics, it is possible that such models may serve to not only describe the characteristics of a research response, but also eventually predict its relative effectiveness in future responses. Applying machine learning and natural language processing approaches to scientific discourse in this way would serve as an important proof of principle for how multidisciplinary research efforts that leverage the expertise of seemingly disparate areas of inquiry (such as bioinformatics, library and information science, academic medicine, and linguistics) can generate innovative solutions to complicated challenges of pandemic proportions.

Disclosure Statement

This research was supported by funding from the Andrew W. Mellon Foundation's Scholarly Communications Program awarded to the Digital Scholarship Center (AWMF 1708-04721).

References

- Baghaei Lakeh, A., & Ghaffarzadegan N. (2017). Global trends and regional variations in studies of HIV/AIDS. *Scientific Reports*, 7(1), Article 4170. <https://doi.org/10.1038/s41598-017-04527-6>
- Bakshy E., Rosenn, I., Marlow C., & Adamic L. (2012, April 16–20). The role of social networks in information diffusion. *Proceedings of the 21st International Conference on the World Wide Web* (pp. 519–528). ACM, <https://doi.org/10.1145/2187836.2187907>
- Bedford, J., Farrar, J., Ihekweazu, C., Kang G., Koopmans, M., & Nkengasong, J. (2019). A new twenty-first century science for effective epidemic response. *Nature*, 575(7781), 130–136. <https://doi.org/10.1038/s41586-019-1717-y>
- Bhimraj, A., Morgan, R. L., Shumaker, A. H., Lavergne, V., Baden, L., Cheng, V.C., Edwards, K. M., Gandhi, R., Muller, W. J., O'Horo, J. C., Shoham, S., Murad, M. H., Mustafa, R. A., Sultan, S., & Falck-Ytter, Y. (2020). *Infectious Diseases Society of America guidelines on the treatment and management of patients with COVID-19*. Infectious Diseases Society of America (IDSA). <https://www.idsociety.org/practice-guideline/covid-19-guideline-treatment-and-management/>
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. <https://doi.org/10.1145/2133806.2133826>
- Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2–3), 143–296. <https://doi.org/10.1561/15000000030>
- Burt, R. (2005). *Brokerage and closure: An introduction to social capital*. Oxford University Press.
- Callaway, E., Cyranoski, D., Mallapaty, S., Stoye, E., & Tollefson, J. (2020, March 18). The coronavirus pandemic in five powerful charts. *Nature*, 579, 482–483. <https://www.nature.com/articles/d41586-020-00758-2>
- Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 288–296). <https://proceedings.neurips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html>
- Croxford, S., Kitching, A., Desai, S., Kall, M., Edelstein, M., Skingsley, A., Burns, F., Copas, A., Brown, A. E., Sullivan, A. K., & Delpech, V. (2017). Mortality and causes of death in people diagnosed with HIV in the era of highly active antiretroviral therapy compared with the general population: An analysis of a national observational cohort. *The Lancet Public Health*, 2(1), E35–E36. [https://doi.org/10.1016/s2468-2667\(16\)30020-2](https://doi.org/10.1016/s2468-2667(16)30020-2)

- Everett, M. G., & Borgatti, S. P. (1999). The centrality of groups and classes. *The Journal of Mathematical Sociology*, 23(3), 181–201. <https://doi.org/10.1080/0022250X.1999.9990219>
- Everett, M. G., & Valente, T. W. (2016). Bridging, brokerage and betweenness. *Social Networks*, 44, 202–208. <https://doi.org/10.1016/j.socnet.2015.09.001>
- Forsythe, S., Stover, J., & Bollinger, L. (2009). The past, present and future of HIV, AIDS and resource allocation. *BMC Public Health*, 9(Suppl. 1), Article S4. <https://doi.org/10.1186/1471-2458-9-S1-S4>
- George, C. P., & Doss, H. (2018). Principled selection of hyperparameters in the latent Dirichlet allocation model. *Journal of Machine Learning Research*, 18(162), 1–38. <https://dl.acm.org/doi/10.5555/3122009.3242019>
- Giamarellos-Bourboulis, E. J., Netea, M. G., Rovina, N., Akinosoglou, K., Antoniadou, A., Antonakos, N., Damoraki, G., Gkavogianni, T., Adami, M. E., Katsaounou, P., Ntaganou, M., Kyriakopoulou, M., Dimopoulos, G., Koutsodimitropoulos, I., Velissaris, D., Koufargyris, P., Karageorgos, A., Katrini, K., Lekakis, V., . . . Koutsoukou, A. (2020). Complex immune dysregulation in COVID-19 patients with severe respiratory failure. *Cell Host & Microbe*, 27(6), 992–1000. <https://doi.org/10.1016/j.chom.2020.04.009>
- Gobat, N., Amuasi, J., Yazdanpanah, Y., Sigfid, L., Davies, H., Byrne, J. P., Carson, G., Butler, C., Nichol, A., & Goossens, H. (2019). Advancing preparedness for clinical research during infectious disease epidemics. *ERJ Open Research*, 5(2), Article 00227-2018. <https://doi.org/10.1183/23120541.00227-2018>
- Gould, R. V., & Fernandez, R. M. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, 19, 89–126. <https://doi.org/10.2307/270949>
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360–1380. <https://doi.org/10.1086/225469>
- Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web* (pp. 491–501). ACM. <https://doi.org/10.1145/988672.988739>
- Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference* (pp. 11–15). SciPy. <https://conference.scipy.org/proceedings/scipy2008/>
- Kapadia, S. (2019, August 19). Evaluate topic models: Latent Dirichlet Allocation (LDA). *Towards Data Science*. <https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0>

- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86. <https://doi.org/10.1214/aoms/1177729694>
- Lee, J. J., & Beckelheimer J. (2020). Anthropocene and empire: Discourse networks of the human record. *Publications of the Modern Language Association*, 135(1), 110–129. <https://doi.org/10.1632/pmla.2020.135.1.110>
- Leung, C. C., Lam, T. H., & Cheng, K. K. (2020). Mass masking in the COVID-19 epidemic: People need guidance. *The Lancet*, 395(10228), 945. [https://doi.org/10.1016/S0140-6736\(20\)30520-1](https://doi.org/10.1016/S0140-6736(20)30520-1)
- Levine C. (2009). Narrative networks. *Novel: A forum on fiction*, 42(3), 517–523. <http://www.jstor.org/stable/27764354>
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)* (pp. 262–272).
- Musial, K., & Juszczyszyn, K. (2009). Properties of bridge nodes in social networks. In N. T. Nguyen, R. Kowalczyk, & S.-M. Chen (Eds.), *ICCI 2009: International Conference on Computational Collective Intelligence* (pp. 357–364). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04441-0_31
- National Institutes of Health. (May 2020). *Translational science spectrum*. <https://ncats.nih.gov/translation/spectrum>
- National Science Foundation. (2016). *Convergent research at NSF*. <https://www.nsf.gov/od/oia/convergence/index.jsp>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830. <http://jmlr.org/papers/v12/pedregosa11a.html>
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining–WSDM '15* (pp. 399–408). ACM. <https://doi.org/10.1145/2684822.2685324>
- Sims, M., & Bamman, D. (2020). Measuring information propagation in literary social networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 642–652). <https://doi.org/10.18653/v1/2020.emnlp-main.47>

- Syed, S., & Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics* (pp. 165–174). IEEE. <https://doi.org/10.1109/dsaa.2017.61>
- Talley, E., Newman, D., Mimno, D., Herr II, B., Wallach, H., Burns, G., Leenders, A., & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, 8(6), 443–444. <https://doi.org/10.1038/nmeth.1619>
- Thorlund K., Dron, L., Park, J., Hsu, G., Forrest, J. I., & Mills, E. J. (2020). A real-time dashboard of clinical trials for COVID-19. *The Lancet: Digital Health*, 2(6), E286–E287. [https://doi.org/10.1016/S2589-7500\(20\)30086-8](https://doi.org/10.1016/S2589-7500(20)30086-8)
- Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K., Ceder, G., & Jain, A. (2019, July 3). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763), 95–98. <https://doi.org/10.1038/s41586-019-1335-8>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., ... Kohlmeier, S. (2020). CORD-19: The Covid-19 open research dataset. *arXiv*. <https://doi.org/10.48550/arXiv.2004.10706>
- Wilkerson, J., Smith, D., & Stramp, D. (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science*, 59(4), 943–956. <https://doi.org/10.1111/ajps.12175>
- World Health Organization. (2020, May 20). *Coronavirus disease (COVID-19)*. Situation Report – 121. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200520-covid-19-sitrep-121.pdf?sfvrsn=c4be2ec6_2
- Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley H. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, 714–715, 1–73. <https://doi.org/10.1016/j.physrep.2017.10.001>

Supplementary Files

The data supporting these findings, as well as all appendices, are available through the [University of Cincinnati Institutional Repository](#). The code for corpus preparation, model generation and network visualization is available at: https://github.com/ucdscenter/Covid_Network_Bridges_code.

©2021 Margaret V. Powers-Fletcher, Erin E. McCabe, Sally Luken, Danny Wu, Philip A. Hagedorn, Ezra Edgerton, Amy Koshoffer, Dorcas Washington, Suraj Kannayyagari, Jason Lee, Jennifer Latessa, Anita Shah, and James Jaehoon Lee. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.