SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search

TOM HOPE, Allen Institute for AI and the University of Washington
JASON PORTENOY* and KISHORE VASAN*, University of Washington
JONATHAN BORCHARDT*, Allen Institute for AI
ERIC HORVITZ, Microsoft Research
DANIEL S. WELD, Allen Institute for AI and the University of Washington
MARTI A. HEARST, University of California, Berkeley
JEVIN WEST, University of Washington

ABSTRACT

The COVID-19 pandemic has sparked unprecedented mobilization of scientists, already generating thousands of new papers that join a litany of previous biomedical work in related areas. This deluge of information makes it hard for researchers to keep track of their own field, let alone explore new directions. Standard search engines are designed primarily for targeted search and are not geared for discovery or making connections that are not obvious from reading individual papers.

In this paper, we present our ongoing work on **SciSight, a novel framework for exploratory search** of COVID-19 research. Based on formative interviews with scientists and a review of existing tools, we build and integrate two key capabilities: first, exploring interactions between biomedical facets (e.g., proteins, genes, drugs, diseases, patient characteristics); and second, discovering **groups** of researchers and how they are connected. We extract entities using a language model pre-trained on several biomedical information extraction tasks, and enrich them with data from the Microsoft Academic Graph (MAG). To find research groups automatically, we use hierarchical clustering with overlap to allow authors, as they do, to belong to multiple groups. Finally, we introduce a novel presentation of these groups based on both topical and social affinities, allowing users to drill down from groups to papers to associations between entities, and update query suggestions on the fly with the goal of facilitating exploratory navigation.

SciSight[1] has thus far served over $10K$ users with over $30K$ page views and 13% returning users. Preliminary user interviews with biomedical researchers suggest that SciSight complements current approaches and helps find new and relevant knowledge.

1 INTRODUCTION

Scientists worldwide are joining forces in an unprecedented concerted effort to understand and treat COVID-19 [2]. Racing against the exponentially growing number of infections, researchers are beginning to make advances: Creating proteins tailor-made to help stop the virus [26], identifying viral genome sequences [19], using artificial intelligence to help pick drug candidates [48], and many more efforts.

However, researchers in biology and medicine have long been wrestling with a very different kind of exponential growth – the flurry of research papers published every year, at a rate that

*Denotes equal contribution
[1]http://scisight.apps.allenai.org/

Authors' addresses: Tom Hope, Allen Institute for AI and the University of Washington, tomh@allenai.org; Jason Portenoy; Kishore Vasan, University of Washington; Jonathan Borchardt, Allen Institute for AI; Eric Horvitz, Microsoft Research; Daniel S. Weld, Allen Institute for AI and the University of Washington; Marti A. Hearst, University of California, Berkeley; Jevin West, University of Washington.

continues to rapidly increase [64]. Thousands of papers just in the last few months have been pouring into the COVID-19 Open Research Dataset (CORD-19) [59]. At the time of this writing, this includes more than 60,000 scientific publications of potential relevance, both historical and cutting-edge, to coronaviruses and other closely related areas in virology, epidemiology, and biology [15, 46, 64]. This growing network of papers contains valuable knowledge for connecting the dots. The goal of this project is to help connect those dots.

The challenge and importance of keeping pace with the growing literature is not unique to the biomedical sciences; it is a growing challenge across all domains of research [34], but it is especially critical in times when new information is both rapidly emerging and urgently needed in short time scales [2, 40].

The predominant way scientists search and consume the literature is through lists of articles in academic search engines [4]. While search engines are a powerful tool for quickly finding documents relevant to a query, they are mostly geared toward targeted search, when the researcher knows what they are looking for. They are less useful for exploration and discovery – finding out the unknown unknowns and making connections that are not obvious from reading individual documents [3, 29, 63]. A recent review of scholarly visualization tools [4] finds that such tools are "applied relatively rarely". This could be partly due to difficulty of designing usable science mapping systems, and therefore a dearth of these applications, but also due to the friction of adopting new information retrieval tools.

To help accelerate scientific discovery, we propose SciSight, a working prototype framework for **exploratory search of the COVID-19 literature**. Unlike many bibliometric tools (see Section 2), we shift the focus from lists of papers—more useful for targeted search—to networks of biomedical concepts and research groups, with the assumption that traversing across concepts and groups better reflects the kind of exploratory search that we are trying to facilitate.

Building search interfaces in science is difficult not only due to the vast complexities of scientific content and language, but also because of the social forces governing this system. Science is a human endeavor, with complex sociological undercurrents that have tremendous effects on the construction of knowledge [30, 32, 41, 49, 56, 62]. Just like in most fields, silos of knowledge exist throughout the biomedical literature [34][2], hindering cross-fertilization between groups and fields, crucial for driving innovation [27, 33]. These silos can have detrimental effects on research advancement that can ultimately impact human lives [40]. These problems are all the more acute when it comes to the COVID-19 pandemic and the information overload that has ensued [9].

The integral role that the social forces play in constructing scientific knowledge is one of the reasons we think it is critical to **leverage the underlying social structure of the research endeavor to bridge groups and disciplines**, potentially facilitating new collaborations and discovery of methods, problems and directions other scientists are working on. We aim to incorporate the underlying social structure into the design interface, to help researchers make connections to other groups, methods and ideas in the literature. Unlike most bibliographic visualizations, we focus on groups and their connections, and integrate this social graph with exploratory faceted search. On a macro level, we hope our approach will help organize COVID-19 research efforts, potentially reducing the amount of redundant work and accelerating collaboration and discovery.

Our main contributions:

- We present SciSight, a working prototype system for exploratory search and visualization of COVID-19 scientific literature and collaboration networks. We construct a novel visualization of research groups and links between them. We employ hierarchical overlapping community detection, relaxing the common assumption in bibliometric co-authorship analysis that

[2]So Long to the Silos, Nature Biotech, https://www.nature.com/articles/nbt.3544

authors belong to one group alone. We present several approaches to build meta-edges capturing topical and social affinities between clusters, and search for groups based on their centrality along both types of edges.
- We review existing work for bibliometric exploration and visualization, including new tools for COVID-19, and discuss the landscape of tools available to researchers. We conduct formative interviews and preliminary user studies with researchers and medical practitioners to discuss their information gathering needs, and find that SciSight is able to complement standard search and help discover new directions and knowledge.
- Finally, we report some initial findings on properties of the research group network, and demonstrate its use in finding potential bridges between communities.

2 RELATED WORK

The field of bibliometric visualization goes back decades [8], with a large and diverse body of work. Recently, there has been a burst of activity in response to COVID-19. In this section, we begin with a review of search and visualization tools sprouting up recently in the midst of COVID-19 literature and then discuss how this research relates to the broader work in this area.

2.1 COVID-19 literature search and visualization tools

In the first few months of the COVID-19 crisis, companies and research institutions released a flurry of literature search tools. The vast majority of them featured academic search interfaces focusing on finding relevant papers. Here, we focus on a subset of prominent tools more closely related to our work, focused on faceted search and visualization.[3]

Faceted search Many of the COVID-19 search tools we reviewed included standard faceted search [23, 54, 67] functionality, enabling users to first submit a query and then filter the retrieved papers according to various facets. In a search tool from Microsoft Azure [42], users search for papers with a standard query (e.g., "covid-19 ebola"), and are able to filter papers by various facets (such as by certain authors or gene mentions). For a given query, semantically similar terms are suggested as additional queries.

In addition to Microsoft's faceted search, other similar services include IBM Watson's COVID-19 Navigator [28], Elsevier's Coronavirus Research Repository [16], the National Institutes of Health's (NIH) LitCovid [43], and Berkeley's Covid Scholar [5]. These search platforms all focus directly on retrieving relevant papers, with lists of paper titles and abstracts as the most prominent and visible feature (see Figure 1). The various services differ mainly by what facets are included. For example, IBM's COVID-19 Navigator allows users to filter papers according to the UMLS biomedical concepts [7] they contain and by publication year. Elsevier's search allows filtering by facets such as high-level paper topics, journals, and authors' organization. The NIH's LitCovid includes as facets the authors' countries, journals and chemicals, and categorizes papers into high-level themes such as *transmission, mechanism* or *treatment*. Finally, Berkeley's Covid Scholar allows for rich query syntax and integrates data from various sources and supports filtering by tags including the NIH's high-level categories.

Concept associations A small number of tools focus on fine-grained concept relations and associations (Figure 2). The semantic visualization tool by Brandies [10] feeds a COVID-19 knowledge graph automatically extracted by the Blender Lab [60] at the University of Illinois (UIUC), into Kibana [44], a data visualization dashboard product. This tool displays interactions between genes, chemicals and diseases in matrix form, with color-coding reflecting occurrence frequency. Accessing papers mentioning two interacting concepts in context is not possible. Word clouds

[3]For a comprehensive list, see https://cord-19.apps.allenai.org/

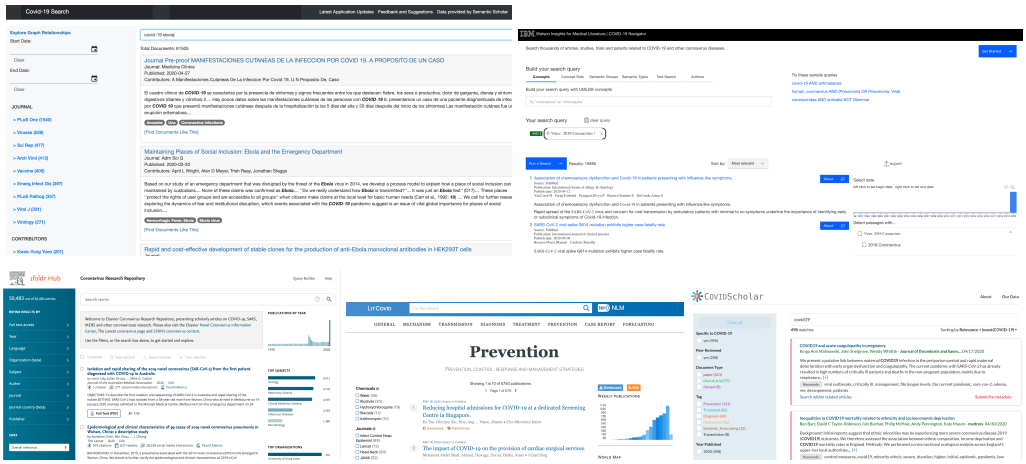


Fig. 1. COVID-19 faceted search interfaces, focusing on papers. From upper left, clockwise: Screenshots of search tools from Microsoft[42], IBM[28], Elsevier[16], NIH[43], Berkeley[5].

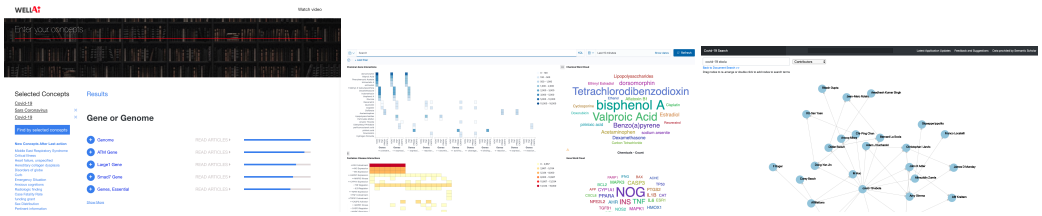


Fig. 2. COVID-19 entity interaction tools. Screenshots from WellAI [61] (left), Brandeis [10] (center) and Microsoft Azure [42] (right).

displaying frequent chemicals and genes are shown, and a structured query search bar based on the Kibana syntax allows users with sufficient system knowledge to query the data. A tool not requiring specialized search syntax is WellAI's [61] exploratory search over biomedical concepts and associations. Once a concept is selected, new suggested concepts appear, along with a bar indicating association strength. Clicking a concept allows users to see a list of relevant papers. A very recent tool from [11] shows clusters of high-level topics extracted with Latent Dirichlet Allocation (LDA) [6], and allows search for sub-topics within clusters and seeing relevant papers.

Author graph visualization In Microsoft Azure's search tool discussed above, users can first enter a search query, and then go to a graph view showing various facets and edges between them. As seen in Figure 2, one graph shows author nodes with directed edges between them, with all edges equally weighted. Double clicking a node adds it as a search term, to further filter papers (which can be viewed upon returning to the document search page).

2.2 Bibliometric visualization

Visualizations of the scientific literature can take many shapes and forms, with the aim of depicting the connections between fields, topics, authors, and, most commonly, papers [4]. While much research has been done in this field over the years, actual tools that are readily available primarily focus on visualization of citation-based graphs between papers, authors or topics [45, 52, 55] and require

Very recently, such tools have seen initial use in visualizing certain aspects of the COVID-19 literature, such as journal networks and heat maps of frequently occurring terms [21]. However, many tools require training before being able to be used, and state of the art bibliometric mapping is currently considered "complex and unwieldy" [4]. Importantly, general adoption by researchers is fairly low, potentially because the typical user "does not *immediately* comprehend a map and (as a result) is not enticed into using it" [13].

In this work, we attempt to improve on current designs for presenting bibliographic information in a search interface. As our key interest is in presenting graphs of research groups in addition to finer-grained entities, we center our review of tools around this area.

Fig. 3. VOSviewer with COVID-19 paper metadata. Node colors represent clusters extracted automatically from the co-authorship graph.

Most bibliometric visualization tools and approaches touch on author relations, and in particular author co-citation or co-authorship network [45, 52, 55, 65]. A nearly ubiquitous feature in these frameworks is the display of individual authors and links between them. While this rich information could in theory be useful, in practice it often renders the visualization inscrutable, especially for real-world networks comprising many authors. Tools in this area often allow users to cluster nodes with community detection algorithms and coloring authors accordingly (see Figure 3), however this typically contributes little in terms of helping users understand and navigate the data. This problem is especially acute when the goal is to enable discovery of new groups and areas, with unfamiliar individual author names.

As a representative example, we briefly review VOSviewer [55], a popular bibliometric visualization tool supporting network visualizations including author clustering. While VOSviewer does not offer an interface dedicated to COVID-19, we download COVID-19 paper metadata from Elsevier's Scopus [16], and then upload to the VOSviewer co-authorship graph visualizer in its supported data format. Figure 3 shows the results. Nodes represent authors, edges represent a co-authorship relation, and colors denote author clusters obtained with community detection (in this case resulting in 13 clusters). While paper metadata was loaded into the tool, the tool did not support searching for specific keywords or author affiliations, or viewing relevant papers and metadata. We observed similar features in other free and commercial tools [45, 52, 65].

Aside from existing tools, a large body of research in human-computer interaction (HCI), information visualization and information retrieval has addressed various problems in the bibliometric sphere, often as a test-case for navigating documents and topics due to their generality beyond the bibliographic domain [14, 20, 29]. In [14] users can gradually explore research areas by viewing papers in the neighborhood of a seed paper, and manually label papers into color-coded groups.

Focusing on searching and visualizing scientific communities, [3] presents the results of PubMed queries as a network diagram, with color-coded nodes representing individual authors, topics, institutions and papers, and investigates the utility of this information with user studies. Studying the history of HCI conferences, [25] presents a large number of bibliometric visualizations including a hybrid matrix network representation [24] for co-authorship groups, showing individual author links within and between each matrix; this work does not display topic labels. In [1], groups of authors are heuristically created based on last authors of papers, and then visualized as special nodes connected to author nodes. Searching for topics and exploring associations is not supported.

Summary In this section, we review bibliometric search and visualization tools and research, including new tools for COVID-19. As expected, we find that a key focus of existing work is on search interfaces for papers, and separately for visualizing graphs of topics, authors, or papers. In the next section, we describe SciSight in light of this work, with the aim of combining faceted navigation and research group detection for exploratory scientific search.

3 SCISIGHT: SYSTEM OVERVIEW

In this section we present an overview of our prototype and its three distinct contributions. We motivate each one by discussing researcher needs that emerged in preliminary formative interviews, in addition to insights from existing systems reviewed in Section 2. We illustrate SciSight's features and exploratory potential with the following illustrative example:

Marc is a researcher interested in exploring areas related to Chloroquine, an anti-malarial drug that has been re-purposed for COVID-19 patients but has been surrounded with various controversies [53]. In particular, Marc wants to find connections between Chloroquine and other drugs and diseases, and to understand how these various entitites are interconnected in order to explore other candidate drugs and potential side-effects. Marc is familiar with the field and its main papers, but the amount of related work is overwhelming with a litany of drugs and diseases. Making things worse, knowing that Chloroquine is not a new type of medication, Marc wants to examine connections made across many years of research and not just recent research.

3.1 Collocation explorer

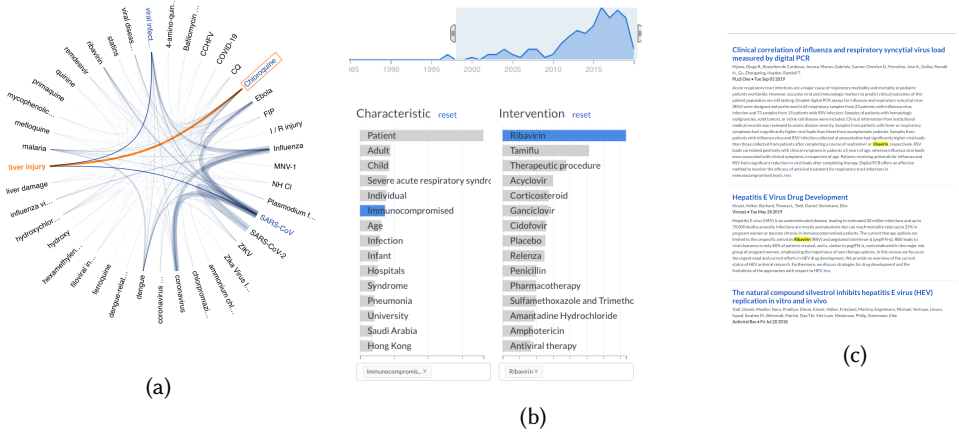
Users of SciSight can search for a term/concept of interest, or get suggestions based on important COVID-19 topics. Searching for a term displays a network of top related terms mined from the corpus, based on term collocation counts across the corpus (co-appearance in the same sentence). Importantly, as seen in Figure 4a, interrelations between all terms are shown (not just with the query), presenting the user with more potential connections to explore. Entities are displayed in a customized chord diagram [35] layout[4], with edge width corresponding to collocation frequency. Upon clicking an edge, a list of papers in which collocations occurred is shown.

Continuing our example, Marc the researcher can search for Chloroquine and see its network of associations, such as its potential connection to liver damage, or its connection to other drugs such as the anti-viral drug Ribavirin. Marc can navigate the graph by clicking nodes to further explore new associations (e.g., clicking liver damage to potentially discover more related drugs and diseases). Navigation is known to help facilitate exploration [29], such as when users do not have a pinpointed query in mind [63].

Entity extraction and selection To extract entities we use S2ORC-BERT [39], a language model pre-trained on a large corpus of scientific papers. This language model is fine-tuned[5] on two separate biomedical named entity recognition (NER) tasks (BC5CDR [37] and JNLPBA [31]), enabling us to extract spans of text corresponding to *proteins, genes, cells, drugs, and diseases* from across the corpus. We extract entities only from titles and abstracts of papers to reduce noise and focus on the more salient entities in each paper. We present only entities collocated at least twice in total with other entities.

[4]Implemented in d3, https://www.d3-graph-gallery.com/chord

[5]See experiment configuration https://github.com/allenai/scibert/blob/master/scripts/exp.py.



(a) (b) (c)

Fig. 4. (a) **Collocation explorer**: corpus-wide associations between biomedical entities, such as drugs and conditions. Highlighted in the figure is the connection between Chloroquine and liver injury. (b) **Exploratory search** of connections between patient characteristics and interventions. Papers and authors working with immunocomprimised patients and Ribavirin would be listed below the facet feature. The graph above shows the number of papers per year with these criteria. (c) **Drilling down**: All our features support drilling down to relevant papers, after the user explores different facets and converges on a query of interest.

Our choice of entities is the result of an initial round of interviews with biomedical experts, identifying these concepts as fundamental to the study of the virus. In particular, we began by launching a probe study focusing on proteins, genes and cellular information and conducting initial interviews with researchers and clinicians (for more details on study participant backgrounds, see Section 4). Participants with a more clinical orientation expressed interest in viewing the associations between drugs and diseases, while users from a biology background wished to focus on proteins, genes and cells. When asked whether they would prefer to have all types of entities in one view, all participants responded with a preference for two separate graphs to avoid clutter and reduce cognitive load.

3.2 Faceted exploratory search

Similarly to other tools, we incorporate a faceted search tool into SciSight. Our focus is on exploration of topics and associations, with relevant papers displayed below the facets for users wishing to dig deeper after refining their search – rather than being featured front and center. When searching for a topic or an author, new suggestions to help refine the search are suggested based on top co-mentions with the initial query to help prevent fixation on an initial topic and boost associative exploration [29]. While scientists are interested in many different types of facets of papers, in our prototype for this feature we aimed at providing one compact set of facets that can cater to a wide range of interests but still be sufficiently granular. Based on formative interviews and a review of biomedical concept taxonomies, we converged on three widely-used topical facets in biomedicine, that capture characteristics of patients or the problem, interventions, and outcomes [49] (see Figure 4b), extracted automatically with text classification trained on annotated biomedical abstracts with distant supervision [57]. In addition, other facets are available such as journal, affiliation and author. The number of relevant publications is presented over time, possibly revealing trends for specific facets. Users can adjust the desired time range and papers and facets displayed update accordingly.

In our future work, we plan to explore giving the user more control on the types of facets displayed.

Having spotted a potential connection to Ribavirin, Marc searches for it under the intervention facet to find out about related patient populations and outcomes, and to see how popular it has been over time (see Figure 4b). A characteristic that pops-up and catches Marc's attention is immunocompromised patients, as he recalls a colleague recently mentioning the risk of treating such populations. He find certain peaks of interest around points in time, and drilling down to papers published around the year 2016 finds a paper with the following conclusion: "No consensus was found regarding the use of oral versus inhaled RBV... such heterogeneity demonstrates the need for further studies ... in immunocompromised hosts." Marc realizes his knowledge of this domain is lacking, and decides to zoom out and find out what groups and labs are working on immunity and viral diseases, perhaps discovering some familiar collaborators.

3.3 Network of science

In the course of our formative studies and interviews, participants were asked about the utility of a tool showing the research areas for different research groups. Participants expressed the need to see what other groups and labs are doing in a convenient and simple manner, in order to help them collaborate, to keep track of competition and to explore new fields. For example, a virologist we interviewed wanted to find new groups that share similar interests to a group of a leading virology researcher, but without direct social links – no overlap of authors. Using this approach, she was able to discover a new approach she was not familiar with (see Section 4). To support such queries and needs, we build a visualization of groups and their ties and integrate this social graph with exploratory faceted search. In our approach explained below, groups of authors are retrieved based on how well they match faceted queries, while giving higher weight to groups with high centrality.

We design our tool with the following components.

3.3.1 Author groups. Unlike most bibliographic visualizations (such as those reviewed in Section 2), we do not show author-level nodes – aiming to reduce visual clutter in terms of the number of nodes and links shown. Instead, as shown in Figure 6, we represent groups of authors as "cards" [22], each card displaying the salient authors, affiliations and topics in that group (with information from Microsoft Academic [50], linked to the papers from the CORD-19 corpus). Cards are color-coded to reflect relevance to the user's initial query – aiming to strike a balance between the relevance and diversity of the results shown. Users may select how many groups to view, zoom in/out, click a group and scroll down to see more detailed information with a full list of the group's topics, authors and papers.

To identify groups, we employ an overlapping community detection algorithm based on ego-splitting [17] so that authors can belong to multiple groups. Our aim is to relax the assumption typically made in bibliometric co-authorship analysis that authors belong to one group alone. In reality, researchers can "wear many hats", and belong to different groups depending on what topic they work on and with whom. In our experiments we focus on authors who have had at least one paper in the CORD dataset since the year 2017, with the aim of exploring groups recently active in this space. We construct a co-authorship network in which links between authors represent collaboration on a publication, weighted by the number of these publications.

We observe that the co-authorship network consists of one giant connected component, and many much smaller separate components (see Table 1). We focus our study here only on the giant component; smaller connected components largely represented author disambiguation errors in the data, or authors who do not have enough of a presence in the literature to have any connections.

Table 1. Co-Authorship Network Statistics. We observe a large network with sparse connections. The giant component has a lower transitivity score but a higher density score than the larger network. <k>, <L> and CC denote average degree, average path length and clustering coefficient, respectively. Density denotes the ratio of existing edges and the number of possible edges in a complete graph; transitivity denotes the ratio existing triangles and possible triads (vertices with two edges).

	Num Nodes	Num Edges	<k>	<L>	CC	Transitivity	Density
Complete Network	86647	521704	12.04	-	0.89	0.65	0.0001
Giant Component	38040	335711	17.65	6.35	0.89	0.598	0.0004

We take the giant connected component of this network (38,040 nodes and 335,711 edges), and run the community detection algorithm. Empirically we observe a small number of "super clusters", large communities with hundreds of authors that appear to be noisy and not densely linked, a well-known characteristic of community structure in real-world networks [36]. We thus apply the clustering algorithm again within any cluster with more than 120 authors to break them down further into more tightly woven groups. This process results in 2,083 clusters. In Figure 5, we show the distribution of cluster sizes. There are 1539 authors belonging to two groups; 2987 are in more than one cluster, and 712 in more than two clusters.

Based on our formative interviews, we know that a primary interest of potential users is a mix of topical and social information – who is working on what and where. To that end, we display the most salient authors (*who*), affiliations (*where*) and topics (*what*). We rank topics by their TF-IDF scores within a cluster. For example, in a cluster where the SARS disease topic is mentioned very frequently, we may down-weight it if it also appears frequently across other groups in the data. Authors and affiliations are simply ranked by relative frequency of appearance in a group.

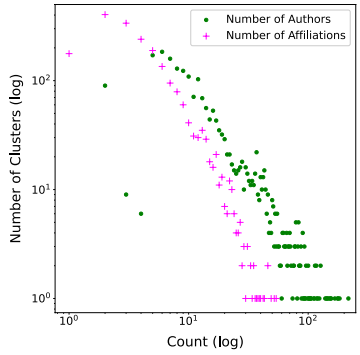


Fig. 5. Two overlaid plots showing the number of authors (green) and number of affiliations (magenta) in clusters, respectively. We observe that a few groups have a large diversity of affiliations.

We allow users to dig deeper into groups with two further levels of resolution. First, when hovering over a group with the cursor, users are shown a tooltip box with the top 5 authors, affiliations and topics, with full names shown. Secondly, upon clicking a group we show full ranked lists of these entities, in addition to the group's papers ranked by recency (with title, abstract, journal and authors, including a hyperlink to read the full paper; see Figure 4c).

3.3.2 Group links. We construct two types of links between groups. The first type (shown as purple edges) represents topical affinity across groups – the interests they have in common based on publishing on similar topics. The second type of link (shown as green edges) captures social affinity between groups, meaning groups with many shared author relationships. By virtue of providing both kinds of links, the tool implicitly provides suggestions for future potential collaborations or connections, particularly when a social connection does not currently exist alongside a topical one.

Cluster Relationships To find the related topics between clusters we try two different approaches, one based on embedding their textual surface form with a language model, and another

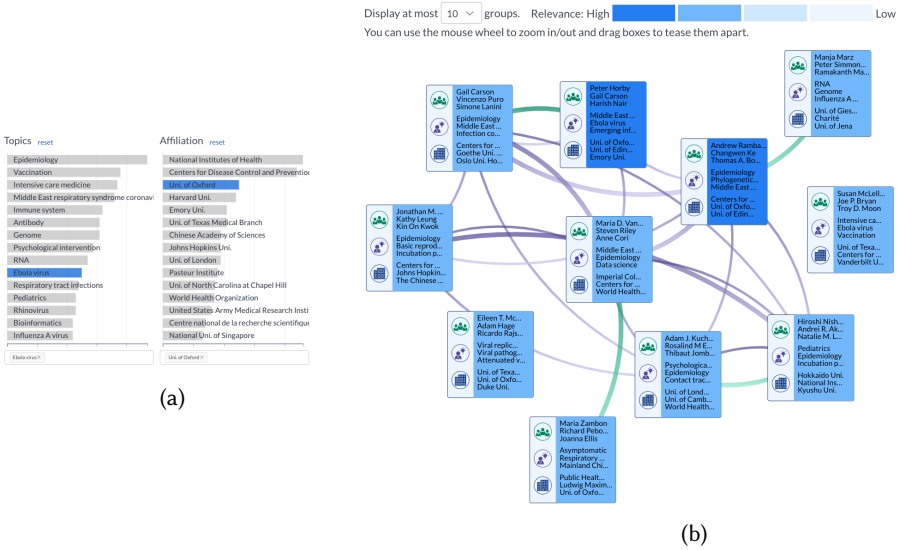


Fig. 6. Exploratory faceted search over clusters of researchers (a) and visualizing the network of groups and their ties with group "cards" (b). Each card has three icons denoting the top three authors, topics and affiliations, respectively. Green edges capture social affinity (shared authors), and purple edges capture topical affinity (interests both groups have in common).

is based on the scores provided by Microsoft Academic's in-house knowledge graph (MAG) [50, 58] of academic entities including topics, authors, venues and papers. We use these similarity scores to discover relationships between topics such as *epidemiology* and *contact tracing*. This allows us to get a better understanding of the focus of various groups.

In the first approach, we use a pre-trained language model trained to capture semantic similarity[6] [47] to get vector representations of the names of topics (such as *contact tracing*). With each topic represented with its embedding, we get a vector representation of groups of authors with a simple weighted average of embeddings as follows. For each paper in a group, we extract its corresponding MAG topics, and treat each group as "bag of topics". We then select the top 10 topics ranked by TF-IDF, embed each topic with the language model, and compute a TF-IDF weighted average to obtain a group-level topic vector.

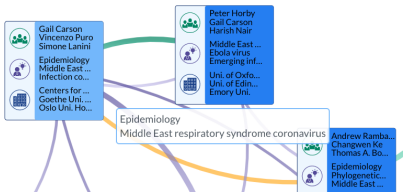


Fig. 7. Link interpretability

We also experiment with a different method, using relatedness scores provided by Microsoft Academic. These scores are found using heterogeneous network embedding models to extract vector representations for nodes (such as topics), allowing to compute relatedness scores. We standardize the scores to be between 0 and 1 and use a TF-IDF weighted average of all pairwise distances between the top 10 cluster topics. Empirically we find in our initial experiments that both approaches perform similarly in terms of finding similar clusters. We leave to future work more rigorous comparisons.

[6]RoBERTa-large[38] fine-tuned on the STS and SNLI benchmarks, github.com/UKPLab/sentence-transformers

Table 2. Example of a bridge we discover: A prominent biologist who is the sole author shared between two research groups, each with different focus and collaborators.

Derek AT Cummings*	Group 1	Group 2
Topics	Medical microbiology Bioinformatics Ebola virus Direct fluorescent antibody test Respiratory virus Emerging infectious disease Polymerase chain reaction Incubation period	Epidemiology Computational epidemiology Contact tracing Data mining Herd immunity Seroprevalence Influenza prevention Attack rate
Authors	Trish M. Perl Nicholas G. Reich Connie S. Price Charlotte A. Gaydos	Kin On Kwok Donald S. Burke Vivian Wan In Wei Isabel Rodriguez-Barraquer
Affiliations	Uni. of Colorado Denver Uni. of Massachusetts Amherst Johns Hopkins Uni. Uni. of Florida	The Chinese Uni. of Hong Kong Uni. of Florida Uni. of Hong Kong Sungkyunkwan Uni.

* - Has appointments to Dept. of Biology, Univ. of Florida and Dept. of Epidemiology, Johns Hopkins

Table 3. Example pairs of clusters our method finds as similar (top row) and dissimilar (bottom). We show the top salient MAG topics for each cluster (ranked by TF-IDF) – the main areas a group focuses on.

Cluster 1 Topics	Cluster 2 Topics	Cosine Similarity
rna, genome, nanopore sequencing, viral evolution	human virome, rna virus, non cellular life, rna silencing	92%
epidemiology, basic reproduction number, contact tracing, branching process	psychological intervention, basic reproduction number, subclinical infection, social distance	89%
food safety, food industry, food processing, hygiene	ards, extracorporeal membrane oxygenation, lung, pulmonary alveolar microlithiasis	11%
disease eradication, disease reservoir	hematopoietic stem cell transplantation, histoplasmosis, histoplasma, lung	11%

Finding Bridges As a test case for demonstrating our framework's ability to find gaps and similarities across groups of researchers, we identify "bridges" between groups, potentially signifying structural holes [12] in the author network. We examine groups that work on *data science* (MAG topic), a highly interdisciplinary field connecting researchers from multiple domains. We discover *Derek AT Cummings*, a prominent biologist and epidemiologist with appointments at two

different universities. We find him to be a sole shared author between two different clusters: one focusing on areas tied with virology and medical microbiology, while the other more associated with computational epidemiology. The former group has 15 authors, and the latter has 35. We show more details about these two clusters in Table 2. Trying to place this author (or others like him) solely in either one of these groups may be difficult, because of his work in multiple areas and with different collaborators. This case study points to our aim of identifying the "multiple hats" of the same author and to make connections across groups and topics.

Similarity Evaluation In a preliminary experiment, we selected 30 random clusters and computed topical affinities to other clusters. For each group we randomly sample one cluster out of the top 3 closely related clusters, and another cluster from the bottom 50% of farthest clusters (for network construction as shown to users, we only create links between top-most similar groups). We randomize the results and give them to a biomedical data analyst for annotation. We find that overall, we are able to correctly find pairs of research groups that work in similar areas with a 80% precision. In our future work we aim to collect validation data enabling to measure both precision and recall, to examine the rate of false negatives (missed similar clusters), too.

3.3.3 Exploratory search. Users can search topics, affiliations, or authors (obtained from Microsoft Academic Graph (MAG) [50] in disambiguated form). We rank topics shown to users based on global TF-IDF scores. As in standard faceted search, queries across facets are conjunctive (e.g., gene sequencing AND University of Washington), and queries within facets are disjunctive (e.g., gene sequencing OR bioassays).

Each query consists of a selection of one or more components in the search interface. This selection automatically reveals new suggested metadata that is frequently associated with the original query, suggesting additional groups and topics to explore. For example, a virologist may be interested in searching for groups connected to the University of Oxford. When searching for Oxford, the Ebola virus topic moves up the list of associated concepts. The scientist can select it to further refine the search. Alternatively, the researcher may decide to pivot away from Oxford by removing that affiliation and shifting their focus around Ebola.

The problem of finding relevant communities to a query has been explored to a certain extent under the rubric of *community search* [18, 51], in which given a graph $\mathcal{G}$ and a set of query nodes in the graph, the objective is to find a subgraph of $\mathcal{G}$ that contains the query nodes and is also densely connected. The problem of community search in heterogeneous networks has only recently been explored [18], and only for one query node. In addition, in our setting we aim to retrieve high-relevance groups, with ranked topics, authors and affiliations. While we leave to future work the development of new approaches for this novel setting, in our preliminary experiments we retrieve relevant results for a user's query with two simple approaches. In the first, we simply compute the overlap between query facets q and the top-K salient facets f for each group of authors, and rank groups by normalized overlap size $\frac{|\{q:q \in f\}|}{|f|}$. In the second approach, we compute weighted PageRank scores [66] over a graph with meta-nodes representing groups of authors, and meta-edges constructed as described earlier in this section. We do so separately for both types of edges: one for topical affinity, and the other for social proximity. At query time, we compute the average of these two scores and the facet overlap score.

4 INFORMAL USER STUDIES AND FINDINGS

We conduct preliminary user studies with four researchers and one practitioner directly involved in COVID-19. One researcher (**P1**) is a research scientist in virology, whose work also studies the Zika virus; a second researcher (**P2**) is a postdoctoral fellow in the area of virology, working on viral infections and human antibody responses, and the third (**P3**) is a postdoctoral fellow and MD

working primarily in Oncology. A fourth participant (**P4**), medical professional and PharmD who has been working with patients and answering pharmacological questions related to COVID-19 treatment that require regular research into both cutting-edge and historical literature. Finally, we interview a researcher (**P5**) working on viral diseases and proteins.

Below we summarize and discuss the main observations and themes that arose in our formative studies. More extensive user studies are planned in the near future.

All participants used the system actively, searching for terms, topics and groups while following a think-aloud protocol in an hour-long session. A member of the research team answered user questions, and asked them to explain their statements (such as why a specific relation or group is interesting to explore). In addition, participants were asked what kind of information they are generally interested in, how they use standard search engines and what issues they face, and for feedback on the utility of certain features (such as group exploration).

Exploratory tools to complement search All users mentioned the need for intuitive tools that can support exploratory needs unanswered by most search engines. **P4** discussed the problem of answering patient questions under uncertainty and lack of familiarity with the emerging COVID-19 knowledge that is constantly evolving. "I am frequently asked my clinical opinion on combinations of medication therapies that, until about three weeks ago, were virtually unheard of." In such cases, questions from healthcare facilities and patients "are often vague, such as 'what are the latest recommendations for medication management?' " As a result, **P4** frequently "needs to look into research", but this is difficult with standard search engines "when you don't know what you don't know." Participants mentioned that viewing associations between fine-grained facets helps them explore in a way that can mitigate these issues ("PubMed search doesn't show associated terms, standard keyword search won't work well for finding related concepts" (**P4**), "nice to have such a tool to complement the usual search engines we use in our field" (**P3**)). **P3**, an MD and researcher, discussed the utility of presenting information "in the form of concepts and links between them" as an "intuitive and convenient way to look into research, that also dovetails with how we think as medical professionals." Users also highlighted the need for a **user-friendly interface**, mentioning that SciSight's representation of groups is intuitive (**P3**), that it has "capabilities beyond the search engine I currently use" (**P4**), and that the "web interface is easier to work with" than other common tools (**P1**), by allowing to explore concepts and dig deeper into papers when relevant.

Finding new groups and directions As part of our formative studies in the process of developing SciSight, we gauged user interest in a visualization that highlights groups of researchers—what they work on and how they are connected— rather than just individual researchers. Participants raised various ways in which they would use such a feature: Finding unknown labs or groups working on similar topics as potential competitors or collaborators (**P2**), exploring around known groups to find related groups and directions (**P2**), understanding what various groups are working on and how relevant they are (**P1, P4**), and unveiling connections to other groups to explore potential conflicts of interest (**P4**). **P1** said this tool would be "useful for the entire field in general," giving an example of a lab [notice group association] that uses assays to identify which proteins antibodies bind to in order to to neutralize HIV, and connecting to other groups working on serum utilization for SARS to potentially collaborate and combine efforts.

To test our tool after our initial round of interviews, we conducted a follow-up study with **P2**. Prior to the session, we asked the participant to think of groups of researchers or areas she'd like to explore. A primary area of focus of **P2** is around HIV; when searching for this topic in SciSight, **P2** found new groups that seemed interesting to explore (based on topics and paper titles) in the first attempt.

Searching for a coronavirus-related gene, **P2** was able to discover several groups of authors and papers that were new: "I was able to find a new perspective on the subject that I would not have

found otherwise". By finding groups that shared strong topical interests but had no strong social overlap, the participant navigated to an unknown group of Chinese authors that published a paper about virus evolution in a local journal **P2** was not aware of. "I would not have come across this journal and paper otherwise, and it's a very different approach. I should definitely read this."

Finally, when searching for a prominent scientist in this field, several connected research groups emerged. Based on **P2**'s interest in epitopes on the SARS-CoV-2 virus Spike protein, the participant found another group associated with the same author, connected via topical affinity with shared focus on epitopes. Selecting the group and examining papers, **P2** found a recent paper [68] with a new direction, and was "curious in to see other related papers in this group."

Discovering unknown associations In the course of **P2**'s exploratory session, **P2** expressed interest in studying in-depth a certain antibody (CR3022) and its associations. With minimal guidance from a member of the research group, the participant searched for the antibody with SciSight's collocation feature, finding "very relevant associations" and also "two potentially surprising and interesting publications. I'm going to look into those papers." **P1** found that that this interactive view of the data "opened up the doors," and **P4** mentioned the tool can "reveal connections that publications are starting to make."

As an example, **P1** searched for cells associated with a type of cytopathic effect and found a specific cell line (Calu-3, a human lung cancer cell line), which led to "spotting an interferon with relevant and interesting studies, very useful." This pattern exemplifies the merit of **associative browsing and discovery**, where navigating around an area of interest in an intuitive manner can lead to new insights [29, 63].

By "starting with keywords which I know are relevant" and exploring associated entities, **P4** "learned that Disulfiram had been studied in-vitro to fight the virus," a connection the participant considered an "unknown unknown". As a related example that demonstrates the potential value of connecting both new and older research, after one round of associative browsing, **P4** discovered a relation between broad-spectrum antivirals and the MERS coronavirus, which provided a "new idea" that is "strongly relevant". **P5**, a viral protein researcher, was able to find specific associations considered new and interesting (such as between the TNF inflammatory cytokine and ERK1/2, a type of protein kinase, considered relevant to **P5**'s interest in cytokine profiles and their correlation with disease severity).

Limitations: more information and features Finally, we also discuss some potential enhancements that emerged as part of our formative studies. Users suggested that user-inputted concepts be collected and combined with existing concepts/terms on-the-fly if they don't yet exist in SciSight (**P1, P4**), and that users should be able to remove edges considered not relevant (**P1**), and collapse/combine synonymous nodes (**P1**). **P3** and **P4** suggested to enable ranking associations by "measures of novelty" to allow users to focus on potentially more emergent knowledge, while **P1** was concerned about losing valuable information from previous studies.

All users expressed interest in seeing many more types of specific entities and also finer-grained relations, but also having the ability to control which ones are shown together. For example, **P4** was interested in seeing relations indicating risk factors, and exploring facets such as metabolizer enzymes, patient weight and metabolic speed, drug dosages and their effects, study sizes, and more. **P3** was more interested in specific viruses, human cells, in-vitro cultures, vaccines, epitopes and mutation machinery, and finding specific techniques and approaches. The diverse interests of experts presents a significant challenge for building systems serving an interdisciplinary user base while maintaining a user-friendly interface, which we plan to explore in our future work.

5 CONCLUSION

In this paper we presented our ongoing work on SciSight, a framework for scientific literature search and exploration. The design of SciSight is informed through a set of formative interviews, as well as a review of existing bibliometric tools. We demonstrate SciSight's use on a large corpus of papers related to the COVID-19 pandemic and previous coronaviruses. Users find that SciSight facilitates connections between related research groups and biomedical concepts in ways different than targeted search.

We use specialized language models to extract fine-grained entities such as proteins, drugs and diseases, and a hierarchical overlapping community detection approach for automatically identifying groups of researchers. Unlike previous work on co-authorship visualization, we display group "cards" rather than individual author nodes, with the aim of providing a more abstract notion of research collaborations (and potentially reducing clutter in terms of the number of nodes shown). We also introduce a novel link scheme capturing topical and social affinities between communities, designed to identify socially disjoint groups working on similar topics. We evaluate our affinity scores with annotations from a biomedical domain expert, finding them to have high precision, and present some initial findings on properties of the research group network.

Preliminary user interviews with scientists and medical professionals suggest that SciSight is able to complement standard search and may pave new research directions. In near-term future work, we plan to conduct extensive user studies with domain experts to validate SciSight and better understand its potential and limitations.

REFERENCES

[1] Moataz Abdelaal, Florian Heimerl, and Steffen Koch. 2017. ColTop: Visual topic-based analysis of scientific community structure. In *2017 International Symposium on Big Data Visual Analytics (BDVA)*. IEEE, 1–8.

[2] Matt Apuzzo and David D. Kirkpatrick. 2020. Covid-19 Changed How the World Does Science, Together. https://www.nytimes.com/2020/04/01/world/europe/coronavirus-science-research-cooperation.html.

[3] Michael E Bales, David R Kaufman, and Stephen B Johnson. 2009. Evaluation of a prototype search and visualization system for exploring scientific communities. In *AMIA Annual Symposium Proceedings*, Vol. 2009. American Medical Informatics Association, 24.

[4] Michael E Bales, Drew N Wright, Peter R Oxley, and Terrie R Wheeler. 2020. Bibliometric Visualization and Analysis Software: State of the Art, Workflows, and Best Practices.

[5] Berkeley. 2020. Covid Scholar. https://covidscholar.org/. last accessed 2020-05-12.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[7] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.

[8] Christine L Borgman and Jonathan Furner. 2002. Scholarly communication and bibliometrics. *Annual review of information science and technology* 36, 1 (2002), 2–72.

[9] Jeffrey Brainard. 2020. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat.

[10] Brandeis. 2020. Semviz, Brandeis. https://www.semviz.org/. last accessed 2020-05-12.

[11] Pierre Le Bras, Azimeh Gharavi, David A. Robb, Ana F. Vidal, Stefano Padilla, and Mike J. Chantler. 2020. Visualising COVID-19 Research. arXiv:2005.06380 [cs.IR]

[12] Ronald S Burt. 2004. Structural holes and good ideas. *American journal of sociology* 110, 2 (2004), 349–399.

[13] RK Buter, ECM Noyons, M Van Mackelenbergh, and T Laine. 2006. Combining concept maps and bibliometric maps: First explorations. *Scientometrics* 66, 2 (2006), 377–387.

[14] Duen Horng Chau, Aniket Kittur, Jason I Hong, and Christos Faloutsos. 2011. Apolo: interactive large graph sensemaking by combining machine learning and visualization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 739–742.

[15] Emmie de Wit, Neeltje van Doremalen, Darryl Falzarano, and Vincent J Munster. 2016. SARS and MERS: recent insights into emerging coronaviruses. *Nature Reviews Microbiology* 14, 8 (2016), 523.

[16] Elsevier. 2020. Elsevier Coronavirus Research Repository. https://coronavirus.1science.com/search. last accessed 2020-05-12.
[17] Alessandro Epasto, Silvio Lattanzi, and Renato Paes Leme. 2017. Ego-splitting framework: From non-overlapping to overlapping clusters. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 145–154.
[18] Yixiang Fang, Yixing Yang, Wenjie Zhang, Xuemin Lin, and Xin Cao. 2020. Effective and efficient community search over large heterogeneous information networks. *Proceedings of the VLDB Endowment* 13, 6 (2020), 854–867.
[19] Peter Forster, Lucy Forster, Colin Renfrew, and Michael Forster. 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences* 117, 17 (2020), 9241–9243.
[20] Carsten Görg, Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo, Haesun Park, and John Stasko. 2012. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Transactions on Visualization and Computer Graphics* 19, 10 (2012), 1646–1663.
[21] Milad Haghani, Michiel CJ Bliemer, Floris Goerlandt, and Jie Li. 2020. The scientific literature on Coronaviruses, COVID-19 and its associated safety-related research dimensions: A scientometric analysis and scoping review. *Safety Science* (2020).
[22] Marti Hearst. 2009. *Search user interfaces*. Cambridge university press.
[23] Marti A Hearst. 2006. Clustering versus faceted categories for information exploration. *Commun. ACM* 49, 4 (2006), 59–61.
[24] Nathalie Henry, Jean-Daniel Fekete, and Michael J McGuffin. 2007. Nodetrix: a hybrid visualization of social networks. *IEEE transactions on visualization and computer graphics* 13, 6 (2007), 1302–1309.
[25] Nathalie Henry, Howard Goodell, Niklas Elmqvist, and Jean-Daniel Fekete. 2007. 20 years of four HCI conferences: A visual exploration. *International Journal of Human-Computer Interaction* 23, 3 (2007), 239–285.
[26] HHMI. 2020. Citizen Scientists Are Helping Researchers Design New Drugs to Combat COVID-19. https://www.hhmi.org/news/citizen-scientists-are-helping-researchers-design-new-drugs-to-combat-covid-19.
[27] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 235–243.
[28] IBM. 2020. Watson Insights for Medical Literature | COVID-19 Navigator. https://covid-19-navigator.mybluemix.net/search. last accessed 2020-05-12.
[29] Sanjay Kairam, Nathalie Henry Riche, Steven Drucker, Roland Fernandez, and Jeffrey Heer. 2015. Refinery: Visual exploration of large, heterogeneous networks through associative browsing. In *Computer graphics forum*, Vol. 34. Wiley Online Library, 301–310.
[30] J Sylvan Katz and Ben R Martin. 1997. What is research collaboration? *Research policy* 26, 1 (1997), 1–18.
[31] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*. Citeseer, 70–75.
[32] Molly M King, Carl T Bergstrom, Shelley J Correll, Jennifer Jacquet, and Jevin D West. 2017. Men set their own cites high: Gender and self-citation across fields and over time. *Socius* 3 (2017), 2378023117738903.
[33] Aniket Kittur, Lixiu Yu, Tom Hope, Joel Chan, Hila Lifshitz-Assaf, Karni Gilon, Felicia Ng, Robert E Kraut, and Dafna Shahaf. 2019. Scaling up analogical innovation with crowds and AI. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1870–1877.
[34] Esther Landhuis. 2016. Scientific literature: information overload. *Nature* 535, 7612 (2016), 457–458.
[35] Sukwon Lee, Sung-Hee Kim, Ya-Hsin Hung, Heidi Lam, Youn-ah Kang, and Ji Soo Yi. 2015. How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 499–508.
[36] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. 2009. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* 6, 1 (2009), 29–123.
[37] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016 (2016).
[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[39] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of ACL*. https://arxiv.org/abs/1911.02782
[40] Michael Loevinsohn, Lyla Mehta, Katie Cuming, Alan Nicol, Oliver Cumming, and Jeroen HJ Ensink. 2015. The cost of a knowledge silo: a systematic re-review of water, sanitation and hygiene interventions. *Health policy and planning* 30, 5 (2015), 660–674.

[41] Göran Melin and Olle Persson. 1996. Studying research collaboration using co-authorships. *Scientometrics* 36, 3 (1996), 363–377.
[42] Microsoft. 2020. Azure Cognitive Search - Covid-19 Search Demo. https://covid19search.azurewebsites.net/. last accessed 2020-05-12.
[43] NIH. 2020. NIH LitCOVID. https://www.ncbi.nlm.nih.gov/research/coronavirus/. last accessed 2020-05-12.
[44] Elastic NV. 2020. Kibana: Your window into the Elastic Stack. https://www.elastic.co/kibana. last accessed 2020-05-12.
[45] Olle Persson, Rickard Danell, and J Wiborg Schneider. 2009. How to use Bibexcel for various types of bibliometric analysis. *Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday* (2009).
[46] Didier Raoult, Alimuddin Zumla, Franco Locatelli, Giuseppe Ippolito, and Guido Kroemer. 2020. Coronavirus infections: Epidemiological, clinical and immunological features and hypotheses. *Cell stress* 4, 4 (2020), 66.
[47] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084
[48] Peter Richardson, Ivan Griffin, Catherine Tucker, Dan Smith, Olly Oechsle, Anne Phelan, and Justin Stebbing. 2020. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *Lancet (London, England)* 395, 10223 (2020), e30.
[49] Connie Schardt, Martha B Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. 2007. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC medical informatics and decision making* 7, 1 (2007), 16.
[50] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*. 243–246.
[51] Mauro Sozio and Aristides Gionis. 2010. The community-search problem and how to plan a successful cocktail party. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 939–948.
[52] Marie B Synnestvedt, Chaomei Chen, and John H Holmes. 2005. CiteSpace II: visualization and knowledge discovery in bibliographic databases. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.
[53] Franck Touret and Xavier de Lamballerie. 2020. Of chloroquine and COVID-19. *Antiviral Research* (2020), 104762.
[54] Daniel Tunkelang. 2009. *Faceted search*. Vol. 5. Morgan & Claypool Publishers.
[55] Nees Van Eck and Ludo Waltman. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* (2010).
[56] Caroline S Wagner and Loet Leydesdorff. 2005. Network structure, self-organization, and the growth of international collaboration in science. *Research policy* 34, 10 (2005), 1608–1618.
[57] Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J Marshall. 2016. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research* (2016). Available from drevidence.com.
[58] Kuansan Wang, Zhihong Shen, Chi-Yuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019. A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data* 2 (2019), 45.
[59] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. CORD-19: The Covid-19 Open Research Dataset. *arXiv preprint arXiv:2004.10706* (2020).
[60] Qingyun Wang, Xuan Wang, Manling Li, Heng Ji, and Jiawei Han. 2020. Knowledge Extraction to Assist Scientific Discovery from Corona Virus Literature. http://blender.cs.illinois.edu/covid19/. last accessed 2020-05-12.
[61] Wellai. 2020. WellAI COVID-19 Machine Learning Analytics For Researchers. https://wellai.health/covid/. last accessed 2020-05-12.
[62] Jevin D West, Jennifer Jacquet, Molly M King, Shelley J Correll, and Carl T Bergstrom. 2013. The role of gender in scholarly authorship. *PloS one* 8, 7 (2013).
[63] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.
[64] Peace Ossom Williamson and Christian IJ Minter. 2019. Exploring PubMed as a reliable resource for scholarly communications services. *Journal of the Medical Library Association: JMLA* 107, 1 (2019), 16.
[65] Pak Chung Wong, Beth Hetzler, Christian Posse, Mark Whiting, Susan Havre, Nick Cramer, Anuj Shah, Mudita Singhal, Alan Turner, and Jim Thomas. 2004. In-spire infovis 2004 contest entry. In *IEEE Symposium on Information Visualization*.
[66] Wenpu Xing and Ali Ghorbani. 2004. Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004*. IEEE, 305–314.

[67] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. 2003. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 401–408.

[68] Meng Yuan, Nicholas C Wu, Xueyong Zhu, Chang-Chun D Lee, Ray TY So, Huibin Lv, Chris KP Mok, and Ian A Wilson. 2020. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* 368, 6491 (2020), 630–633.