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

Tom Hope[♣,♡] **Jason Portenoy**[♣,♥*] **Kishore Vasan**[♥*] **Jonathan Borchardt**[♣*]

Eric Horvitz[♠] **Daniel S. Weld**[♣,♡] **Marti A. Hearst**[◇] **Jevin West**[♥]

[♣]Allen Institute for Artificial Intelligence

[♡]Paul G. Allen School for Computer Science & Engineering, University of Washington

[♥]Information School, University of Washington

[♠]Microsoft Research [◇]University of California, Berkeley

{tomh,jasonp,jonathanb}@allenai.org {kishorev,jevinw}@uw.edu

Abstract

The COVID-19 pandemic has sparked unprecedented mobilization of scientists, generating a deluge of papers that makes it hard for researchers to keep track and explore new directions. Search engines are designed for targeted queries, not for discovery of connections across a corpus. In this paper, we present **SciSight, a system for exploratory search** of COVID-19 research integrating two key capabilities: first, exploring associations between biomedical facets automatically extracted from papers (e.g., genes, drugs, diseases, patient outcomes); second, combining textual and network information to search and visualize *groups* of researchers and their ties. SciSight[1] has so far served over $15K$ users with over $42K$ page views and 13% returns.

1 Introduction

Scientists worldwide are racing against the growing number of COVID-19 infections, to understand and treat the disease (Apuzzo and Kirkpatrick, 2020). However, a very different kind of exponential growth has been plaguing researchers – the flurry of papers published every year, at a rate that continues to increase (Williamson and Minter, 2019). At the time of this writing, the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020a) includes over 130,000 publications of potential relevance, both historical and cutting-edge.

To boost scientific discovery over this corpus, we propose SciSight, a working prototype system for **exploratory search of the COVID-19 literature**. Unlike many tools (see Section 2), we shift the focus from searching over lists of papers or authors, to navigating **networks of biomedical concepts and research groups** – for example, exploring links between COVID-19 and other diseases, or labs working on treatments. While search engines are a powerful tool for finding documents, they are mostly geared toward targeted search, when researchers know what they are looking for – less useful for exploring connections that are not obvious from reading individual papers (Bales et al., 2009; White and Roth, 2009).

Building exploratory interfaces in science is difficult not only due to the complexities of scientific content, but also because of social undercurrents that have tremendous effects on the construction of knowledge (Wagner and Leydesdorff, 2005; Pan et al., 2012; West et al., 2013) (as reflected, for example, in biased citation patterns (King et al., 2017)). Silos of knowledge throughout the literature (Vilhena et al., 2014)[2] can hinder research advancement and cross-fertilization across groups and fields that is crucial for driving innovation (Hope et al., 2017; Kittur et al., 2019), ultimately impacting human lives (Loevinsohn et al., 2015). These problems are acute when it comes to the COVID-19 pandemic, with new information rapidly emerging and urgently needed.

We aim to incorporate the social structure into an intuitive design interface, to help researchers **make connections to other groups and ideas in the literature** by traversing across networks of concepts and groups of scientists – helping users discover **who is working on what, and where?**

We identify groups by clustering the co-authorship graph, and extract topics and entities from the group's papers. Each group is represented using textual and network information: the group's salient authors (*who*), the topics they

[*] Equal contribution.

[1] http://scisight.apps.allenai.org/

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

work on (*what*) and their affiliations (*where*). We build meta-edges capturing topical affinity between groups using a language model fine-tuned for semantic similarity, and present approaches for searching for groups with queries consisting of authors, topics and affiliations. Each selected query automatically suggests new queries to try, to support exploration (Kairam et al., 2015).

In summary, our main contributions:

- A working prototype for exploratory search and visualization of COVID-19 scientific literature and collaboration networks, based on a fusion of automatically extracted textual information (topics, entities) and co-authorship network information.
- User interviews with experts suggest that SciSight can help complement standard search and help discover new directions.

2 Related work

The field of bibliometric visualization goes back decades (Borgman and Furner, 2002), with a large body of work. 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 (Bales et al., 2020). 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 *individual* authors, papers or topics (Van Eck and Waltman, 2010; Synnestvedt et al., 2005; Persson et al., 2009). 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. This problem is especially acute when the goal is to enable discovery of new areas with unfamiliar authors.

Recently, such tools have been applied to COVID-19 papers, such as journal networks and heat maps of frequently occurring terms (Haghani et al., 2020). However, many tools require training before being able to be used, and state of the art bibliometric mapping is currently considered "complex and unwieldy" (Bales et al., 2020), potentially because the typical user "does not *immediately* comprehend a map and (as a result) is not enticed into using it" (Buter et al., 2006).

COVID-19 tools. In response to COVID-19, many tools for exploring the relevant literature have been released. The great majority featured paper search interfaces, with lists of titles and abstracts being the main focus. Many of the COVID-19 tools we reviewed included standard faceted search functionality (Yee et al., 2003; Hearst, 2006; Tunkelang, 2009), enabling users to filter papers according to various facets. In a search tool from Microsoft Azure (Microsoft, 2020), for example, users can filter search results by various facets (such as by authors or gene mentions extracted automatically from texts). Similar services were made available by IBM Watson (IBM, 2020), Elsevier (Elsevier, 2020) and the National Institutes of Health (NIH, 2020).

Currently a small number of tools focus on concept associations. One tool (Tu et al., 2020) feeds a COVID-19 knowledge graph (KG) from (Wang et al., 2020b) into Kibana[3], an external product for creating dashboards with complex heat maps of term frequencies in documents, including a specialized query language for users with sufficient familiarity with Kibana. A tool from (Bras et al., 2020) shows clusters of high-level topics extracted with Latent Dirichlet Allocation (LDA) (Blei et al., 2003) (visualized with word clouds).

In this paper, we integrate textual information from papers and the network of author collaborations, allowing users to drill down from research groups to papers to associations between entities in one system, with a custom interface aimed to help users "comprehend the map" (Bales et al., 2020; Buter et al., 2006) intuitively.

3 SciSight: system overview

In this section we present an overview of our prototype and its distinct components. We motivate each by discussing researcher needs. We illustrate SciSight's features and potential with the following illustrative example:

Marc is a researcher interested in exploring Chloroquine, an anti-malarial drug that has been surrounded with controversies in the context of COVID-19 (Touret and de Lamballerie, 2020). In particular, Marc wants to find connections between Chloroquine and other drugs and diseases, and to understand how these entities 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 re-

[3] https://www.elastic.co/kibana

lated work is overwhelming with a litany of drugs and diseases. Complicating things further, knowing that Chloroquine is not a new medication, Marc wants to examine connections across years of research, not just recent work.

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). Entities are displayed in a customized chord diagram (Lee et al., 2015) layout[4], with edge width corresponding to collocation frequency. As seen in Figure 1a, interrelations between all terms are shown (not just with the query), presenting the user with more potential connections to explore (users can also control the number of entities shown). Clicking an edge between two entities displays a list of papers containing both terms.

Continuing our example, Marc can search for Chloroquine and see its network of associations, such as a 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 (Kairam et al., 2015), such as when users do not have a pinpointed query in mind (White and Roth, 2009).

Entity extraction and selection To extract entities we use S2ORC-BERT (Lo et al., 2020), a new language model pre-trained on a large corpus of scientific papers. This model is fine-tuned[5] on two separate biomedical named entity recognition (NER) tasks (BC5CDR (Li et al., 2016) and JNLPBA (Kim et al., 2004)), 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 show only entities collocated at least twice with other enti-

[4]SciSight is implemented with React, server-side Crossfilter, DC.js, D3.js, and Varnish.

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

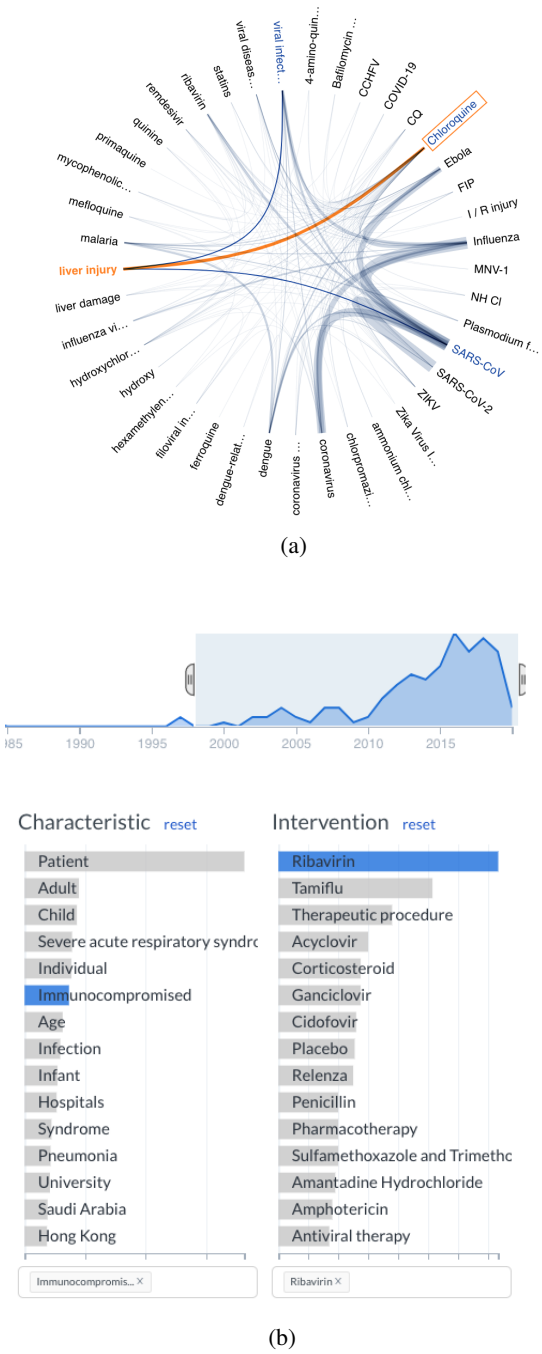

Figure 1: (a) **Collocation explorer**: corpus-wide associations between biomedical entities, such as drugs and conditions. Highlighted in the figure is the edge between Chloroquine and liver injury. (b) **Exploratory search** of connections between patient characteristics and interventions. Papers working with immunocomprimised patients and Ribavirin would be listed below the facet feature. The time graph above shows the number of papers per year with these criteria.

ties. 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. Participants with a more clinical orientation expressed interest in viewing 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, participants responded with a preference for separate graphs to avoid clutter and reduce cognitive effort.

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 presented based on top co-mentions with the initial query to help prevent fixation on an initial topic and boost associative exploration (Kairam et al., 2015). 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 (Schardt et al., 2007) (see Figure 1b), extracted automatically from biomedical abstracts with the distant supervision model in (Wallace et al., 2016). In addition, CORD-19 metadata facets are available, such as journal, affiliation and author. The number of relevant publications is shown over time, possibly revealing trends for specific facets. Users can adjust the time range to update the papers and 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 often it has been mentioned over time (see Figure 1b). A characteristic that pops-up and catches Marc's attention is immunocompromised patients, as he recalls a colleague discussing the risk of treating such populations. He finds peaks of interest around some points in time, and drilling down to papers from around 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 also discovering some familiar collaborators.

3.3 Network of science

In the course of formative interviews with domain experts, participants expressed the need to see what other groups are doing in order to keep track, explore new fields and potentially collaborate. We build a visualization of groups and their ties and integrate this social graph with exploratory faceted search over topics, authors and affiliations. We design our tool with the following components.

3.3.1 Author groups

To identify groups of researchers, we start by constructing a co-authorship network in which links between authors represent collaboration on a paper, weighted by the number of papers. We then employ an overlapping community detection algorithm based on ego-splitting (Epasto et al., 2017) so that authors can belong to multiple clusters (groups). We relax the assumption typically made in 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 they work on and with whom.

As shown in Figure 2, we represent groups with "cards" (Bota et al., 2016) of salient authors, affiliations and topics (with information from Microsoft Academic Graph (MAG) (Sinha et al., 2015)). 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 to see a list of its topics, authors and papers.

To explore groups recently active in this space we select authors with at least one paper in CORD-19 since the year 2017. We focus on the giant connected component of this network (111,236 author nodes, 951,072 edges; smaller components typically represented disambiguation errors), and run the community detection algorithm. We observe a small number of "super clusters", large communities with hundreds of authors not densely linked, a

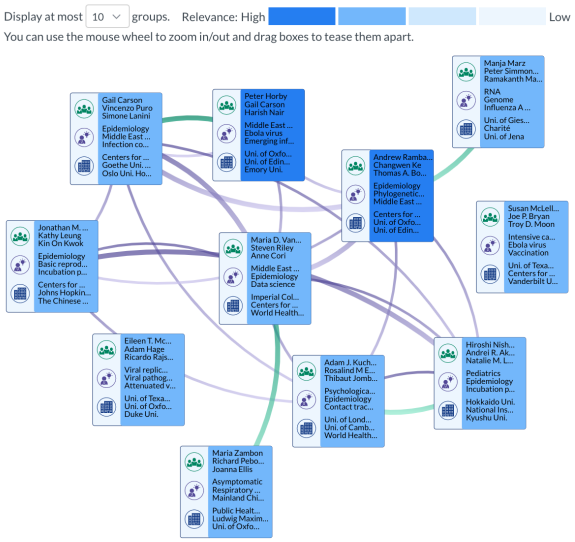

Figure 2:
Visualizing the network of groups with group "cards". Each card has three icons denoting the top three authors, topics and affiliations, respectively. Card color indicates relevance to the search query. Green edges capture social affinity (shared authors), and purple edges capture topical affinity.

well-known characteristic of community structure in real-world networks (Leskovec et al., 2009). We thus apply the clustering algorithm again within clusters with more than 120 authors to break them down further into denser groups. This results in 6,475 clusters. There are 5276 authors belonging to two groups; 6657 are in more than one cluster, and 1381 in more than two clusters.

We display a mix of textual and social information: the most salient authors (*who*), affiliations (*where*) and topics (*what*). We rank topics by their TF-IDF scores within a cluster, and authors and affiliations by relative frequency in a group. Users can also 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).

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 providing both kinds of links, the tool implicitly suggests potential collaborations or connections, particularly when a social connection does not currently exist alongside a topical one.

Cluster Relationships To find topical affinity between clusters we embed topic surface forms with a language model trained to capture semantic similarity[6]. With each topic represented with its embedding, we get a vector representation of groups with a simple TF-IDF weighted average of embeddings, and compute cosine similarity.

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 (Burt, 2004) 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.

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 future work we plan to collect validation data enabling to measure both precision and recall.

3.3.3 Exploratory search for groups

Users can search topics, affiliations, or authors. We rank topics based on global TF-IDF scores. As in standard faceted search, queries across facets

[6] RoBERTa-large-STS-SNLI(Liu et al., 2019) github.com/UKPLab/sentence-transformers

are conjunctive (e.g., gene sequencing AND Harvard), and queries within facets are disjunctive (e.g., gene sequencing OR bioassays). Each query consists of one or more choices for each facets. A selection automatically suggests new facets that are frequently associated with the original query, suggesting more groups and topics to explore.

The problem of finding relevant communities to a query has been explored to a certain extent under the rubric of *community search* (Sozio and Gionis, 2010; Fang et al., 2020), 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 (Fang et al., 2020), and only for one query node. In addition, in our setting we aim to retrieve high-relevance groups, with ranked topics, authors and affiliations. We propose to retrieve relevant results for a user's query with two simple approaches. In the first, we compute the overlap between query facets q and the top-K salient facets f for each group, and rank groups by normalized overlap size $\frac{|\{q:q\in f\}|}{|f|}$. In the second approach, we compute weighted PageRank scores (Xing and Ghorbani, 2004) over a graph with meta-nodes representing groups, and meta-edges 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 scores and the facet overlap score.

4 Informal user studies and findings

We conduct preliminary user studies with four researchers and one practitioner. **P1** is a research scientist in virology, whose work also studies the Zika virus; **P2** is a postdoctoral fellow in the area of virology, working on viral infections and human antibody responses, and **P3** is a postdoctoral fellow and MD working primarily in Oncology. **P4** is a medical professional and PharmD. **P5** is a researcher working on viral diseases and proteins.

Discovering unknown associations. Based on interest in a the CR3022 antibody , **P2** searched for it 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** searched for cells linked to a type of cytopathic effect and found Calu-3 (a human lung cancer cell line), which led to "spotting an interferon with relevant and interesting studies, very useful." **P4** discovered a link between broad-spectrum antivirals and MERS, providing a "new and strongly relevant idea". **P5** was able to find associations considered new and interesting between the TNF inflammatory cytokine and ERK1/2, a type of protein kinase (relevant to **P5**'s interest in cytokine profiles).

Finding new groups and directions. **P1** gave an example of a group (lab) using 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. When searching for groups tied with a prominent scientist, **P2** found a group associated with the scientist with recent work (Yuan et al., 2020) revealing a new direction regarding SARS-CoV-2 epitopes. **P2** also found a group in China with shared focus on epitopes but no social ties, revealing "new perspectives that I would not have found otherwise" on virus evolution.

Limitations: more information and features. **P1, P4** suggested that user-inputted concepts be combined with existing concepts/terms on-the-fly, and that edges could be removed by the user. **P3** and **P4** suggested ranking associations by "measures of novelty" to allow users to focus on emergent knowledge. Participants mentioned a diverse range of other entities to explore, e.g., patient weight, metabolic speed, drug dosages, vaccines, mutation mechanisms, and various techniques.

5 Conclusion

In this paper we presented SciSight, an evolving system for scientific literature search and exploration. We demonstrate SciSight's use on a large corpus of papers related to the COVID-19 pandemic and previous coronaviruses. We use state-of-the-art scientific language models to extract entities such as proteins, drugs and diseases, and an overlapping community detection approach for identifying groups of researchers. To visualize groups we display group "cards" with a novel link scheme capturing topical and social affinities between communities, designed to identify socially disjoint groups working on similar topics. Preliminary user interviews suggest that SciSight can help complement standard search and may pave new research directions. In future work, we plan to conduct extensive studies to validate SciSight and better understand its potential and limitations.

References

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.

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*, volume 2009, page 24. American Medical Informatics Association.

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.

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

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

Horatiu Bota, Ke Zhou, and Joemon M Jose. 2016. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 131–140.

Pierre Le Bras, Azimeh Gharavi, David A. Robb, Ana F. Vidal, Stefano Padilla, and Mike J. Chantler. 2020. Visualising covid-19 research.

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

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

Elsevier. 2020. Elsevier coronavirus research repository. https://coronavirus.1science.com/search. Last accessed 2020-05-12.

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*, pages 145–154.

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):854–867.

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*.

Marti A Hearst. 2006. Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49(4):59–61.

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*, pages 235–243.

IBM. 2020. Watson insights for medical literature — covid-19 navigator. https://covid-19-navigator.mybluemix.net/search. Last accessed 2020-05-12.

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*, volume 34, pages 301–310. Wiley Online Library.

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*, pages 70–75. Citeseer.

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*, 3:2378023117738903.

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):1870–1877.

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):499–508.

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):29–123.

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.

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*.

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*.

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):660–674.

Microsoft. 2020. Azure cognitive search - covid-19 search demo. https://covid19search.azurewebsites.net/. Last accessed 2020-05-12.

NIH. 2020. Nih litcovid. https://www.ncbi.nlm.nih.gov/research/coronavirus/. Last accessed 2020-05-12.

Raj Kumar Pan, Kimmo Kaski, and Santo Fortunato. 2012. World citation and collaboration networks: uncovering the role of geography in science. *Scientific reports*, 2:902.

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*.

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):16.

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*, pages 243–246.

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*, pages 939–948.

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.

Franck Touret and Xavier de Lamballerie. 2020. Of chloroquine and covid-19. *Antiviral Research*, page 104762.

Jingxuan Tu, Marc Verhagen, Brent Cochran, and James Pustejovsky. 2020. Exploration and discovery of the covid-19 literature through semantic visualization. *arXiv preprint arXiv:2007.01800*.

Daniel Tunkelang. 2009. *Faceted search*, volume 5. Morgan & Claypool Publishers.

Nees Van Eck and Ludo Waltman. 2010. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*.

Daril A Vilhena, Jacob G Foster, Martin Rosvall, Jevin D West, James Evans, and Carl T Bergstrom. 2014. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1:221.

Caroline S Wagner and Loet Leydesdorff. 2005. Network structure, self-organization, and the growth of international collaboration in science. *Research policy*, 34(10):1608–1618.

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*. Available from drevidence.com.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020a. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*.

Qingyun Wang, Xuan Wang, Manling Li, Heng Ji, and Jiawei Han. 2020b. Knowledge extraction to assist scientific discovery from corona virus literature. http://blender.cs.illinois.edu/covid19/. Last accessed 2020-05-12.

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).

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):1–98.

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):16.

Wenpu Xing and Ali Ghorbani. 2004. Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314. IEEE.

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*, pages 401–408.

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):630–633.