Harnessing Scholarly Literature as Data to Curate, Explore, and Evaluate Scientific Research

Jason Portenoy

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Jevin D. West, Chair

Emma Stuart Spiro

William Gregory Howe

Program Authorized to Offer Degree:
Information School

University of Washington

Abstract

Harnessing Scholarly Literature as Data to Curate, Explore, and Evaluate Scientific Research

Jason Portenoy

Chair of the Supervisory Committee:
Associate Professor Jevin D. West
Information School

There currently exist hundreds of millions of scientific publications, with more being created at an ever-increasing rate. This is leading to information overload: the scale and complexity of this body of knowledge is increasing well beyond the capacity of any individual to make sense of it all, overwhelming traditional, manual methods of curation and synthesis. At the same time, the availability of this literature and surrounding metadata in structured, digital form, along with the proliferation of computing power and techniques to take advantage of large-scale and complex data, represents an opportunity to develop new tools and techniques to help people make connections, synthesize, and pose new hypotheses.

This dissertation consists of several contributions of data, methods, and tools aimed at addressing information overload in science. My central contribution to this space is Autoreview, a framework for building and evaluating systems to automatically select relevant publications for literature reviews, starting from small sets of seed papers. These automated methods have the potential to help researchers save time and effort when keeping up with relevant literature, as well as surfacing papers that more manual methods may miss. I show that this approach can work to recommend relevant literature, and can also be used to systematically compare different features used in the recommendations.

I also present the design, implementation, and evaluation of several visualization tools. One of these is an animated network visualization showing the influence of a scholar over

time. Another is SciSight, an interactive system for recommending new authors and research by finding similarities along different dimensions. Additionally, I discuss the current state of available scholarly data sets; my work curating, linking, and building upon these data sets; and methods I developed to scale graph clustering techniques to very large networks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I thank Jevin West for his guidance, support, and encouragement throughout my entire graduate school journey, from being introduced to this whole "data science" thing, to putting the final touches on this dissertation. I especially thank Jevin for connecting me to so many different people, and always sparking enthusiasm for new collaboration and innovation among everyone involved. I thank my committee—Emma Spiro, Bill Howe, and Mako Hill—for their insight and expertise, and for striking the right balance between supporting and challenging me. I thank the other mentors and collaborators who have helped me along the way, including Bryna Hazelton, Ariel Rokem, Paul Bennett, Ryen White, Eric Horvitz, Dan Weld, Tom Hope, Marissa Radensky, Jessica Hullman, Megan Finn, and Chirag Shah. I thank my funders throughout my time as a graduate student, including JSTOR, Science History Institute, NSF, National Academy of Sciences, Military Suicide Research Consortium, Center for an Informed Public, Microsoft, CZI, and AI2. Thanks to David McDonald, whose one-on-one Python lessons were invaluable at the beginning as I gained the skills needed to do this work. Thanks to my cohort, all my fellow PhD students, and the entire iSchool community, for creating a supportive and inclusive place to earn my doctorate. Thank you to all of my friends and family. And finally, a huge thank you to my parents Drs. Susan Sussmann and Russell Portenoy, my siblings Matthew and Allison, and my girlfriend Jenni Whitney. Your unwavering support and pride over these years has made an enormous difference.

```
while True == True:
    for person in [mom, dad, matthew, allison, jenni] + friends + family + committee + mentors:
        thank_you(person)
```

Chapter 1

INTRODUCTION

Science is a massively parallel endeavor to build upon human knowledge. As the scale and complexity of this body of knowledge increases well beyond the capacity of any individual to make sense of it all, the need for new tools and techniques to help make connections, synthesize, and initiate new questions becomes more pronounced. The high-level goal of my research is to make steps toward **harnessing scientific output as data** in order to address the overwhelming volume and complexity of scientific research and the need to synthesize this material. This dissertation comprises several contributions of data, methods, and tools. First, I describe lessons I have learned from working with a variety of data sets, as well as code I have written to help make use of it. I also present a new method for scaling network clustering to very large citation networks. Second, I develop methods for automating literature review—one of the more important curatorial activities in science for addressing an expanding literature. Using the citation graph and text content of millions of research papers, I develop methods for automatically recommending collections of papers relevant to a given topic or field. Finally, I present visualization tools I have designed and developed to explore and evaluate these collections and other aspects of the data.

The rapid expansion of the literature, in the context of new technology, has a historical parallel in the mid-17th century. In that time period, the first scientific journals were introduced, heralding a major revolution in the production and development of science. It was a new technology—the printing press—that made this revolution possible, and the move from a letter-writing system to one of collected and published research brought with it a major shift in the way science was disseminated and evaluated. The increased volume of scientific output led, by necessity, to new systems of evaluation and curation, such as editorial review

and peer review. We are currently in the middle of a similar technology-driven revolution in science—these technologies being the internet, digital publication, and social media [148]. Once again, the new technologies are driving explosive growth in the amount of available information.

It was estimated in 2014 that there were at least 114 million English-language scholarly documents available on the public web, with tens of thousands more being added every day [95, 199]. While it is increasingly difficult to define and measure the scope of scientific output, recent (2020) estimates count more than 389 million articles indexed by Google Scholar [78]. The rate at which new articles are being added is increasing as well [25], and this exponential growth shows no signs of slowing (see Figure 2.1). Furthermore, the metadata associated with these documents—including authors, keywords, journals, and the relations between all of these entities represented by, for example, citations and coauthorship—make this a highly complex system. This information is increasingly accessible thanks to the internet and advances in digital computation and storage. Taken together, we can view all of this as a body of data with high volume, velocity, and variety. These are the core characteristics originally used to describe the concept of "Big Data"[155], and indeed the term "Big Scholarly Data" has gained some traction in describing this domain [197, 200].

It is in this context that my research sits. The scale of the scientific literature overwhelms traditional, manual methods of curation and synthesis, and so there is a need for innovative new approaches to assist in this endeavor. At the same time, the availability of this literature and surrounding metadata in structured, digital form, along with the proliferation of computing power and techniques to take advantage of large-scale and complex data, represents an opportunity to make meaningful progress in this space. This is what I mean by "harnessing science as data". Treating scholarly literature as large-scale structured data affords us the opportunity to ease the burden on scholars to contextualize their research, and to aid in knowledge building.

The COVID-19 pandemic provides an example of why this matters, and why the SciSight (chapter 4.3) project I am working on has gained some attention. The Semantic Scholar team

at the Allen Institute for Artificial Intelligence (AI2), in collaboration with the White House Office of Science and Technology Policy and several other research institutions, has been collecting the scientific papers relating to the disease and the novel coronavirus in a dataset called the CORD-19 corpus [185]. The current version of CORD-19 contains hundreds of thousands of papers, including almost 200,000 papers published in 2020, and is being updated daily. To put it another way, a new paper is being added to this corpus on average once every two minutes. It is not possible for researchers to keep up with this volume and velocity of new information. A number of tools have emerged to help address this problem [29, 88]. Many of these tools are search-based, serving the needs of users who know what they are looking for. On the other hand, SciSight, and many of the other tools I have worked on, are meant for exploratory search [12, 195], allowing people to discover new connections and patterns in the literature. Chapter 4 has more on SciSight and these other interactive tools.

With the overwhelming scale of the scientific literature as context, the central question this dissertation addresses is: **How can the scholarly literature data be leveraged to curate, explore, and evaluate scientific research?** I will offer several contributions of data, methods, and tools to work toward answering this question. The central methods contribution of my dissertation is the Autoreview project, which offers a framework for developing and evaluating automated methods to generate literature reviews—collections of papers important to a topic—starting from a set of seed papers. Other projects serve as support for this. These other projects include my work curating, linking, and analyzing large bibliometric data sets; and a suite of interactive visualizations and tools I have developed to explore collections of papers, including those collections generated by Autoreview.

1.1 Science of Science

My dissertation fits fairly well within the burgeoning academic community of science of science. The science of science is an emerging field of research that is still in the early stages of defining itself. Its rise, which began in earnest during the first several years of my graduate studies, can be seen in a surge of new research and interest, including publications in and

special issues of high-profile journals like Science [8, 69, 170], interest from funding agencies like the NSF,[1] and popular new conferences like the 2019 Metascience Symposium. The field owes its existence in large part to the field of *scientometrics*, a sub-field of bibliometrics. Scientometrics, since the foundational works of Price [161] and Garfield [72] in the middle of the 20th century, has sought to quantitatively measure and characterize the scientific literature. Science of science is in some ways a rebranding of this field in the wake of the big data revolution—in fact, a parallel can be drawn between the relationship between "science of science" and "scientometrics," and that of "data science" and "statistics" (see [55] for a discussion of the relationship of these latter two terms).[2]

Whereas scientometrics has been largely concerned with measuring various aspects of the scientific literature, science of science has a more ambitious scope, seeking to model literature, researchers, institutions, ideas and other entities, as well as the relationships between them, in order to understand and predict the progress of science at a high level. Two recent papers have sought to define and contextualize the field. Fortunato et al. characterize the goal of the field to be "a quantitative understanding of the genesis of scientific discovery, creativity, and practice and developing tools and policies aimed at accelerating scientific progress." They see the rise of the field as being driven by two key factors: the availability of large scale scholarly data, and collaborations among researchers from different backgrounds and with different skills. They identify as a goal of the field the development of "tools and policies that have the potential to accelerate science." [69] I am well positioned to make contributions to this field—I am part of an information school, an interdisciplinary department with ties to data science, machine learning, sociology of science, and human-computer interaction; I have developed expertise in working with large data sets around scholarly output and collaboration; and my

[1] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505730

[2] The origins of science of science actually go back much further to at least the early 20th century writings of J.D. Bernal [19]. This history is outlined in the introduction to [102]. This introduction includes a footnote nodding to the new brand of "science of science" which is my focus. One interesting aspect about this history is that although scientometrics has its roots in Bernal and the original science of science, it moved away from studying science as a social process. The new brand of science of science, [102] notes, is moving back toward this view.

research is aimed at helping to accelerate science by assisting researchers in finding relevant literature, and identifying and bridging knowledge gaps in science.

Zeng et al. identify a similar goal for science of science: "to understand, quantify and predict scientific research and the resulting outcomes." They specifically trace the field as a development of complex systems research, owing to the rise of the field of network science starting in the late 1990s. They identify a number of specific topics within the purview of science of science, including evaluating papers and scientists, understanding and modeling structural and dynamical patterns, predicting the evolution of a system, identifying paths to success in science, and studying the creation and diffusion of knowledge [207]. This survey largely discusses the field's potential to study broad patterns and trends in scientific research, taking it as a given that these things are worth studying. While my work often focuses on practical applications to help provide insight to individual researchers, I do agree that these high-level patterns in the process of science are interesting in and of themselves, and some of my work does explore these questions. My recent work on SciSight (page 86), for example, has a component that studies the connections and gaps between different researchers working on a broad range of problems across computer science. Another research project in collaboration with sociologists Katherine Stovel and Lanu Kim, recently published in JASIST, is a high-level science of science study of the role that journals play in the success of papers also published as preprints [97].[3]

It is worth noting the difference between research that seeks to describe existing phenomena in science, and research that aims to find ways to intervene and accelerate science. Much of the research in scientometrics and science of science falls in the former category. However, [69] point toward the potential of the latter kind of research, discussing "tools . . . to accelerate science," as well as research on "the integration of machine learning and artificial intelligence in a way that involves machines and minds working together." Traditional scientometric research has actually had a profound, not entirely positive, impact on the process of science.

[3]While I made substantial contributions to this work, it is outside the scope of the dissertation and so will not be included as a chapter.

The development of quantitative measures of impact and success for papers, journals, and authors has shaped scientists' incentives and changed behaviors over time [5, 60, 65, 66]. However, these effects are not the primary aim of this type of research, and in most cases they are not intentional. Research that incorporates methods and lessons from human-computer interaction (HCI), and machine learning and artificial intelligence, on the other hand, is explicitly intended to intervene in science by helping scientists do their work. With this intent in mind, science of science research can be more mindful of the impact it has on science, and aim toward more positive outcomes. Much of my work falls into this category of creating tools to help facilitate scientific progress.

These surveys focus on networks and complex systems as being the major perspective of the emerging field, especially in terms of methods. However, there is also some work being done around natural language processing (NLP), quantitatively studying the *content* of the scientific literature, as opposed to the structure. Combining analyses of both content and structure has potential for many of the goals of science of science [194, 198, 211]. Interest in this potential can be seen in the growing workshop series on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) [39], and in several recent special issues of journals [6, 112]. My own experiences interning at Meta and Semantic Scholar have been illustrative of the promise of joining these worlds together. The research at both of these organizations is largely focused on using NLP to assist researchers in navigating the literature. In these internships, I have brought a network science perspective to join with these methods, studying how to use author collaboration networks to draw attention to interesting connections between groups of researchers. Science is a social process, and these social networks can complement what NLP techniques may miss as they focus primarily on the text content of papers.

1.2 Research and Projects

The central methods contribution of my dissertation is the Autoreview project (chapter 3), which lays out a framework for a big data approach to building and evaluating systems to

generate references for literature reviews. The project has also included implementations of this framework, and demonstrations of its utility both quantitatively and qualitatively.

Bookending this central chapter are chapters that support the Autoreview work. The first of these (chapter 2) will detail the work I have done in curating, linking, and analyzing the scholarly data sets on which Autoreview depends. This work has required a considerable investment of time and effort, but it is necessary as it is a foundation for all the other work that comes after it. In building new tools and approaches from scratch, I have gained deep knowledge of the positives and negatives of different scholarly data sets, and developed methods to combine them to meet different needs. This chapter also includes contributions of methods and tools in my work scaling community detection algorithms to analyze very large citation networks (section 2.2). The other chapter (chapter 4) is on tools that I have built to visualize and explore collections of papers, such as the collections generated by Autoreview. This also includes my current ongoing work on SciSight, an interactive tool for facilitating discovery and finding gaps in networks of science.

Chapter 2

SCHOLARLY PUBLICATION DATA AND CITATION NETWORKS

This chapter discusses the large-scale scientific publication data sets that I have been analyzing and maintaining, and the work I have done using parallel computing and cloud resources to perform clustering on very large citation networks. This can be seen as "backend" work—it is not directly visible in the applications that come in the following chapters, but it serves as a foundation to support it. These efforts do not lead directly to research publications. However, they are critical to the rest of my research, and they can also benefit other researchers working on related problems.

2.1 Data Sets

A recurring maxim in discussions of data science is that its practitioners spend 80% of their time cleaning data [147]. The things about the modern deluge of big scholarly data that make it so promising—its large scale and multifacetedness—are also the things that present significant challenges to its use. Much of my effort has been in understanding the nuances and potential of different available data sets, and curating and linking these data sets to facilitate goals further down the pipeline. This work has been critical as a foundation both for my own research and projects, and for others who want to incorporate bibliometric and science of science analysis into their work. The breadth of data with which I have worked has given me an understanding of both the possibilities of using and linking all of this data for research, as well as the problems that can be hidden within. As part of my contribution, I leave behind the lessons I have learned, as well as some of the code I have developed to work with these data sets.

In this section, I discuss some of the data sets I have worked with, including their affordances and challenges. The two most important are the Web of Science and the Microsoft Academic Graph. These are two of the largest, most comprehensive available bibliographic data sets, allowing for large-scale analyses of science. I also discuss some of the smaller and auxiliary data sets I have worked with, and my experience linking them together.

2.1.1 Web of Science

The **Web of Science (WoS)** was the historic first data set of citation indices [72], and remains one of the most valuable resources for doing research on the science of science. The version of this data set I currently maintain in the Datalab comprises 73 million articles and 1.2 billion citations, and is receiving regular updates. A major challenge with the use of this data set lies in its *ownership*—the Web of Science is the property of Clarivate Analytics (formerly a part of Thomson Reuters), and this company serves as the gatekeeper for how much access we as researchers have to the data. Researchers with access to a WoS subscription (e.g., through a university library) are able to access data for individual records through a web portal. Researchers or groups can also purchase access to an API to download bulk data. In the Datalab, we actually have a more privileged position: we are granted access to the full database underlying WoS, whereas most researchers are limited to what can be downloaded by the WoS API. A recent review of the science of science literature [207] identified the largest subset of WoS that has been analyzed by researchers as being around 47 million papers and 526 million citations—about half as many as in the database I maintain. Despite our access to the WoS data, however, the terms of our agreement with Clarivate limit certain public-facing applications. Another challenge with the WoS data set is around *author disambiguation*—for analyses and applications that study research on the level of individual authors, WoS lags in terms of easily identifying these authors and linking them to their articles. Still, it is a reliable source of relational data for publications across the full scope of science, and I have made use of it in several projects. In my paper introducing Autoreview (see page 31), for example, I used WoS to collect the reference lists from hundreds of review articles in order to

implement the framework, evaluating the task of recommending the most relevant papers from millions of candidate papers.

2.1.2 Microsoft Academic Graph

Microsoft Academic Graph (MAG) is a scholarly data set powered by Microsoft's Bing web indexer [160]. In contrast to WoS, the data collection behind MAG is much more hands-off, using automated web-crawlers and machine learning to infer academic entities (papers, authors, journals/conferences, etc.) and relationships from mostly online content. This approach results in a more comprehensive data set than WoS, with many more articles included (see Figure 2.1). MAG is similar to Google Scholar (see below), with an important difference being that Microsoft makes its data available for researchers to a far greater extent than Google. The team behind MAG is also an active participant in open research: they have recently published open-access research papers about MAG and associated projects [183, 184]. MAG tends to have more noise than WoS, including duplicate entities and missing or incorrect data, but can be more reliable for author and affiliation data. It has been a challenge to maintain and build upon this data set, which contains almost 250 million publications and more than 1.6 billion citations, and is continually updated, but it has paid off in many ways. For instance, in the SciSight project (see page 86), I have used the MAG data to construct co-authorship networks, which I have clustered in order to identify related groups of researchers working on COVID-19 research. I have done community detection on the full citation graph (see next section on Network Analysis), which I have used in models for automated literature review, as well as in visualization and analysis of overlapping research domains (see the section on the cluster comparison network visualization to explore the intersection of InfoSec and Ethics, page 83). On top of this, my expertise with the data has attracted the attention of people who have come to me with requests to help them use the data for their own purposes. These include organizations such as the National Academy of Sciences and HICSS, and individuals such as graduate students and researchers within and outside UW. One example of this is the global publications dashboard developed by the

University of Washington's Office of Global Affairs, which uses MAG data I prepared to show research collaborations UW faculty have had around the world.[1]

In May, 2021—while this section of this dissertation was still being written—Microsoft announced that it would be discontinuing everything under its Microsoft Academic umbrella, including MAG.[2] This means that, after 2021, MAG will no longer be updated, nor accessible through Microsoft (although snapshots of the data will be retained by the Datalab and other organizations and individuals who have made use of it). The rationale Microsoft gives for this decision is that it wants to expand its mission "to have intelligent agents gather knowledge and empower humans to gain deeper insights and make better decisions" beyond academia to enterprise and education. This is a disappointing loss for the science of science community and everyone who has benefited from this data. Microsoft's contribution to open data and collaboration in this space has been admirable, but its statement that the Microsoft Research project "has achieved its objective to remove the data access barriers for our research colleagues" seems dubious, as many of these colleagues rely on data that will be very difficult to recreate without the help of Microsoft's resources. Still, despite this setback, the future still looks bright overall. Microsoft's statement claims that "the momentum is gaining on an open and community-driven alternative . . . "—and while it doesn't give any specifics on this, it does point out a number of other similar resources available to the community, several of which I discuss below. See section 2.1.7 for more discussion on what the future of scholarly data might look like.

2.1.3 Other comprehensive scholarly data sets

In choosing which source of data to use for a particular project or research question, it is important to consider a number of factors. Some of the most important of these factors relate to the data's cost and terms of use. It is unfortunate that these considerations must

[1] https://www.washington.edu/global/publications/

[2] https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/

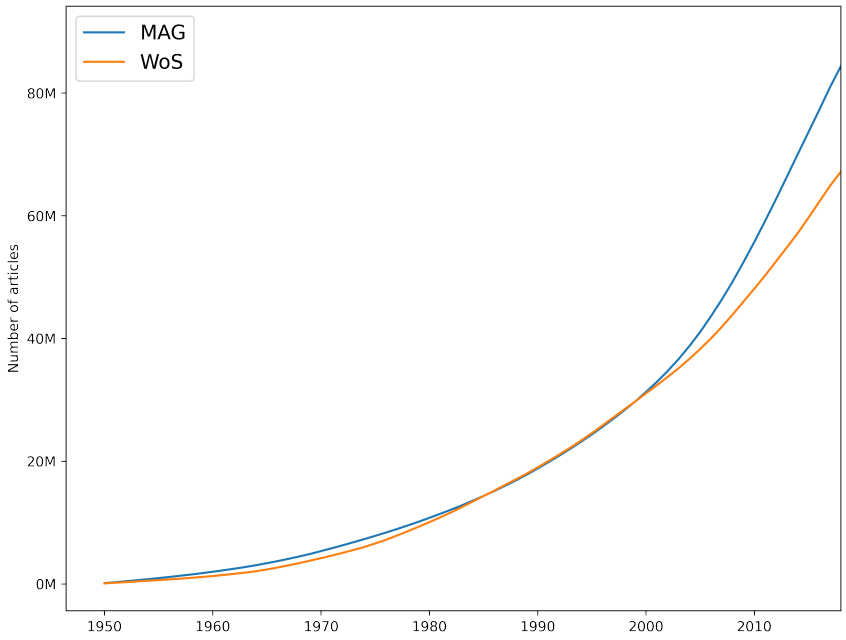

Figure 2.1: Growth of the scientific literature over time. For MAG, journal and conference articles and book chapters are included; patents, repositories, and data sets are excluded.

figure so prominently, as they have little to do with what is best for a given research question or application. Nevertheless, they tend to have a big impact on the decision. Even in the case where the research question is about comparing different data sources, these issues come into play. A recent large-scale comparison of some of the major bibliographic data sources illustrates this point. The authors of this study emphasize the limitation that their analysis of Web of Science is incomplete, because their lab does not have full access to the data. They even tailor their methods around these constraints, stating, "We use Scopus rather than WoS as the baseline because we do not have access to the full WoS database. Our use of Scopus as the baseline does not mean that we consider Scopus to be our preferred bibliographic data source" [176]. The situation is similar in UW's Datalab, except that we have access to the full WoS data, and tend to rely on it rather than other similar data sources like Scopus.

Scopus is considered to be one of the two most established comprehensive data sources for bibliometrics, the other being Web of Science [176]. Scopus emphasizes the size and comprehensiveness of its data, while WoS is more selective, filtering out what it considers articles of lower quality. Scopus is owned by Elsevier, which is also a scientific publisher. This is a potential conflict of interest which could affect how Scopus selects and presents its data; however, they allege that all content selection is carried out by an external and independent group of subject matter experts called the Content Selection and Advisory Board [9].

Besides Web Of Science, Scopus, and Microsoft Academic Graph, there exist a few, relatively newer, comprehensive sources for scientometric data. Dimensions [82] is a new service owned and licensed by Digital Science that links articles and citations to other data such as grants, clinical trials, and altmetric data (such as tweets and blog posts). Crossref is a DOI registration service that also freely provides citation data made public thanks to the Initiative for Open Citations, or I4OC. Elsevier has been a notable holdout in making citation data public, which limits the power of Crossref's data [176]. While I have not explored these data sources yet, I look forward to doing so in the future and seeing how they fit into the evolving landscape of scholarly data sets.

Google Scholar, the academic search engine and database built on the web indexing model and data of Google, is likely the most comprehensive source of bibliometric data in existence [176]. Because Google Scholar lacks a publicly accessible API or any other means of collecting bulk data, using it as a data source for large-scale science of science studies is extremely challenging. One such recent study [110], for example, required a research team to work for three months collecting data that they estimate would take one or two days to collect if a public API were available, according to a member of the team [61]. This is a frustrating situation, as Google Scholar has widespread usage among researchers, and contains valuable information that isn't available elsewhere [52]. In a 2014 interview, one of the co-creators of Google Scholar shot down the idea of having a public API, stating, "Our indexing arrangements with publishers preclude it." This answer is somewhat unsatisfying, seeing as other scholarly search services have found ways to make their data more accessible while maintaining relationships with publishers. Ironically, in this interview, the Google Scholar co-creator identifies a feature he would like to see in the future: the ability "to get you the articles that you need, but that you don't know to search for" [172]. This is a core research focus of myself and others, and our efforts would be significantly aided if Google Scholar were to make their data more open.

2.1.4 Other data sources

In addition to the two main sources of scholarly metadata above, I have worked with a variety of other data sets. ArXiv, an electronic preprint service for scientific articles, has hundreds of thousands of full-text articles and metadata available to download. As part of a science of science study examining how preprint publication correlates with impact, I linked the arXiv data to MAG, which allowed us to track articles' accumulation of citations from preprint, through journal publication, and beyond (see section 2.1.5 for more) [97]. I have also made use of arXiv's full text data in machine-readable LaTeXformat in a project analyzing the use of mathematical symbols in publications across different fields [192]. JSTOR, an online archive of journal articles, has partnered with the Datalab to make use of not only publication data,

but usage data as well. This has made possible some interesting analysis of the distributions of views and downloads of articles, which has complemented a similar analysis of citation patterns. Finally, my internships with Meta and Semantic Scholar have led me to work extensively with their in-house data.

2.1.5 Building on top of existing data sources

Most often, the data sources available to use for my projects and analyses are only a starting point. As part of my work, I have come up with various tools and methods to clean, manipulate and combine data. I offer some examples of this work here, as well as code examples that I make available for others to adapt.

Seeing as different sources of data have different strengths and limitations, it is sometimes useful to combine multiple data sources to answer certain questions or achieve certain ends. One example of this is in a project with the National Academy of Sciences in which I analyzed and recommended papers on science communication and misinformation (see section 4.2). I was initially provided a set of papers from Web of Science. In order to show the people working in these domains, I created a visualization of the authors of these papers along with their co-authorship relationships. Because Web of Science does not have reliable disambiguated author data, I linked the data set to Microsoft Academic Graph. For some papers, a common Document Object Identifier (DOI) could be found in both WoS and MAG, but the DOI was missing for either or both in a substantial number of cases. Paper title is the second most reliable identifier across these data sources, but exact string matching of titles often does not work, due to small variations in the title data. To solve this, I came up with a method to do fuzzy string matching between the data sources, using Elasticsearch and Levenshtein distance. I make the code I developed for this available;[3] this code allows for the comparison of two sets of documents, using approximate matching of titles to find confident matches

[3]https://github.com/h1-the-swan/doc_titlematch

between the sets. With this method, I was able to find confident matches for most of the papers, which I then linked to MAG's author data to construct a co-authorship network.

For the project analyzing citations of articles pre- and post journal publication [97], linking two data sources—MAG and arXiv—was a central part of the research. The goal was to identify papers that had been published to arXiv as preprints and had later been accepted for journal publication, and to analyze incoming citations to these papers both before and after journal publication. Using this analysis, we could separate a paper's citation-based influence from the "journal effect" of it being published in a particular venue. This allowed us to address the research question: does publishing in a high-status journal make any difference for a paper with regards to the citations it receives? The data from arXiv provided a set of papers which had been published as preprints, and the data from MAG provided citation information both before and after journal publication. To link these data sources, I used a similar method as above, identifying common DOIs first and then using a fuzzy matching strategy on titles (see Appendix S1 of [97] for more details). Any incoming citation to one of these papers that was before the paper's journal publication date could be assumed to be a citation to the preprint version of the paper. With this data set, we were able to determine that publication in a high-status journal does indeed predict a higher number of citations for articles (although there is an increasing trend, in the subfields we studied, that authors of high-quality preprints choose never to publish in journals at all).

Working with these large data sets requires careful consideration of how to store and how to programmatically interact with the data. Because of this, the data formats used for storage and access make a difference, and so I have done some work building tools to help with data conversion. For my first few years of doing this kind of research, I used MySQL databases to store and organize data from MAG, WoS, and other sources. However, I have come to decide that for most of my use cases, it is most convenient to store the data as flat files—using the compressed columnar Apache Parquet format—and load it all into memory as needed. This is made possible due to modern hardware on which I can make use of 100GB or more of RAM for extended periods of time. In some cases, the data processing tasks exceed

the amount of RAM available; the Parquet format is useful here as well because it allows for easy use in the Apache Spark parallelization tool.

I have built tools to convert the raw scholarly data for further processing. For example, the data from MAG comes in large dumps of text—CSV files. The code I have published automatically validates and converts these files to Parquet format, to be used later either by the in-memory analysis library Pandas, or the parallelized analysis library Spark.[4] I have also written code to convert network edgelist files—such as paper citation data provided by MAG—to Pajek format, which must be done for certain applications like Infomap clustering (section 2.2).[5]

2.1.6 Data limitations

The research and applications I work on rely heavily on the data sources mentioned above, and while this means they are able to leverage the power of these data, it also means that they inherit certain limitations and issues that come with the data. Some of the work described in this chapter is meant to mitigate these problems. Through experience, I have come to understand ways to clean the data for different downstream tasks. However, this can only go so far, and many problems inherent in the data will inevitably end up affecting any application in which they are used.

The way in which data quality affects a final product—research or other application—depends in part on the scale. My projects that use author-level data, for example the nautilus diagram (section 4.1) and SciSight (section 4.3), can be dramatically affected by these issues. A key culprit here is author disambiguation [162], in which the authors in the data are not properly assigned to their publications. This results in authors with missing papers, or authors being given credit for papers they did not write. The problem is exacerbated with common author names. As discussed above, the degree of disambiguation problems varies across data sets; WoS does not have reliably disambiguated authors, for example, while MAG

[4]https://github.com/h1-the-swan/mag_csv_to_parquet

[5]https://github.com/h1-the-swan/pajek_tools

uses automated methods to provide disambiguated data. All data sets, however, suffer from these issues to some degree. For many bibliometric studies, it is important to track the full career and research portfolio of an author. Given the millions of authors to track, it is infeasible to hand-check every author, which is why most rely on automation. However, even the best automated approaches are error prone. The inevitable errors in disambiguation can diminish overall trust, which complicates user evaluation of these projects.

Authors are not the only entities difficult to disambiguate. Institutions can have similar data issues, although this is a somewhat more tractable problem than for authors due to the smaller number of institutions and the fact that they tend to be documented in more formal ways. Efforts such as GRID (Global Research Identifier Database), a free and openly available data source for disambiguated data on research organizations, are very helpful for doing institutional-level science of science research.[6] Publications are another type of entity that can have disambiguation issues—this happens when multiple versions of a publication exist online or in print. This has become an increasing challenge with the rise of preprint publishing, a trend which has seen considerable growth in recent years, and especially explosive growth around COVID-19 research [30]. In some cases, a research document can change significantly after initial preprint publication, and while preprint services generally do a good job of documenting multiple versions (e.g., arXiv), resolving these entities in aggregate publication data sets is not always a straightforward task (see section 2.1.7 for more discussion on this).

While disambiguation issues can be glaring when working at the scale of individual papers or authors, they tend to be less noticeable in macro-scale research projects that look at authors or papers in aggregate. When analyzing all of the papers in a given year, for example, duplicate or missing papers will be much less obvious than when looking at a single author's papers. These more hidden errors are less of a concern if the data quality problems are the result of random noise, but if they reflect systematic biases then they can be even more insidious. For example, older papers could have more missing citations than more recent

[6] https://grid.ac/

papers, because it is more difficult to track citations for papers published before digital publication was commonplace. This could affect studies which include temporal analyses by artificially inflating citation counts for more recent papers. As another example, author disambiguation errors tend to be more common for Chinese and Korean names, many of which tend to be shared among more researchers and can be anglicized in different ways; any cross cultural analysis must be very careful about these biases [162]. Disambiguation errors are also likelier to occur when authors change their names, a situation which will disproportionately affect representation for women and trans scholars.

Another limitation inherent in the citation networks that I often use are the coarse-grained nature of the data: the citation links between publications are binary in that they either exist or do not, and are not further characterized by, for example, sentiment or importance, or by citation type (e.g., basis, comparison, use, etc.) [53, 91]. I often use these citations as a proxy for influence or impact between publications or groups of publications; this assumes that all citations from a given publication are equal, and that they exist because the cited work had a (generally positive) influence on the citing work. These assumptions do not always hold true. Authors can cite work for a number of different reasons. In a given research paper, some cited work may have a very strong influence, while other citations may be more perfunctory and less influential [171, 209]. Some citations may indeed be negative, in cases where the citing work is criticizing or refuting previous work [1, 7]. In many aggregate bibliometric analyses, the coarse-grained assumptions which ignore these differences likely do not have a huge effect on the conclusions one can draw from the data. However, one can assume that science of science would be generally improved by incorporating more of this nuanced information. This is an active area of research, and improvements in this area will benefit researchers and practitioners who make use of the data.

Data coverage is another limitation. Coverage, in terms of articles, journals and conferences, citations, and other data types, varies across different data sets. I have discussed some of these differences in the above sections—for example, WoS is rather selective in its coverage, preferring to include only articles it considers as higher-quality, while MAG's coverage is much

broader. It must be kept in mind that coverage, or lack thereof, is usually not equal within data sets. Often, certain fields of study are underrepresented in the data, typically Social Sciences and Arts and Humanities. Natural Sciences, Engineering, and Biomedicine tend to have better coverage. The data sets also tend to be heavily biased toward English-language journals and conferences [118]. These characteristics of the data sets must be carefully considered when doing any sort of analysis, but especially cross-cultural or cross-disciplinary analysis.

2.1.7 What comes next? The future of scholarly data

In much the same way that science itself is built on incremental progress and gradual consensus, the available scholarly data is on a journey, with researchers and practitioners working toward making it more and more useful in the endeavors of studying science and assisting scholars. Although more attention often goes to the tools and analyses that are developed, the underlying data is critical for any work in this area, and a focus must be maintained on improving it. Ideally, we could have something akin to a "universal" data set that can represent all of science. This data set would reflect a broad consensus on the scope and representation of the data, and the community would commit to improving and maintaining its quality and accessibility. Several broad issues to keep in mind while working toward this are those of openness, transparency, and data quality.

In general, science of science research and the scientific enterprise as a whole benefits from a culture of open information and accessible data. Despite ongoing issues around ownership and gatekeeping, there are several encouraging trends that point to continuing progress toward this goal. Among the most promising signs is the Initiative for Open Citations (I4OC), a collaboration between publishers, researchers, and others with the goal of making scholarly citation data available in standardized formats and without restrictions. Founded in 2017, this initiative has gained momentum in the form of support from a large number of publishers and partner organizations. Elsevier has been a notable holdout in allowing access to their citation data, but there is pressure for them to make concessions (see for example the recent

protest resignation of the editorial board of Elsevier's *Journal of Informetrics*). Since the initiative's founding, the fraction of publications with open references has increased from 1% to 87%, according to their website.[7] I4OC also has a sister initiative, the Initiative for Open Abstracts (I4OA), which has a parallel goal of making abstracts openly available and machine-readable. These initiatives are part of a larger movement of Open Science, which has among its goals the full access by everyone to both full-text research and surrounding metadata. As progress is made toward this goal, universities and organizations such as Microsoft, AI2, and Meta can build more useful technology that leverages the data to help researchers and decrease friction in the scientific process.

A separate issue is one of consensus and standards for defining what exactly are articles, versions, citations, etc. Currently, the various bibliometric databases all have slightly different perspectives on these matters. Web of Science is selective in its definition of which articles should be included, and Scopus, Dimensions, and Microsoft Academic try to be more comprehensive [176]. Dimensions explicitly lays this out as part of their philosophy: "The database should not be selective but rather should be open to encompassing all scholarly content ... The community should then be able to choose the filter that they wish to apply to explore the data according to their use case" [82]. This is a great approach in theory, but in practice it relies heavily on transparency in terms of which records are being considered and which are excluded. This transparency must be shared among the data providers, the users of the data, and the final product. There needs to be, among all stakeholders, a common understanding of exactly what data is being provided, and how it is being filtered to address specific research questions or applications. Currently, this line of transparency and common understanding is often lacking. This is not entirely surprising given how new these data sets are and how quickly they are growing. It is important that the appropriate attention be given to these issues and that they be addressed as time goes on.

---

[7] https://i4oc.org

Apart from these issues of openness and transparency, the advancement of the field hinges on the quality of the data. Problems here can include missing data, incorrect or duplicate entries, and improperly disambiguated author data. Having accurate, clean, and machine-readable data is key for researchers and practitioners to be able to make the most progress. More nuanced issues exist as well, for example the role of predatory journals that may not contain legitimate scholarly content. The role of preprint articles is another gray area in the world of scientific publication—here there are issues around how these preprints should be included in the data, and how multiple versions of the same document should be resolved and represented. These data issues intersect with the issues of consensus and transparency above, especially since the role of preprints and alternative models of peer review is currently a rapidly evolving facet of scientific publication. Overall, there remains a lot of work to be done in the face of these challenges. We must maintain our focus on the data in order to advance the field. However, in my experience, it seems that significant progress is being made, and we have reason to expect a bright future for scholarly data.

2.2 Network Analysis

A key aspect of large-scale scholarly publication data is the relationship structure that exists between the different papers, authors, and other entities. Understanding the patterns in these relations can help in navigating the literature. This way of looking at the data is the basis for the complex network approach to the science of science, with these entities comprising vertices and the different types of relations between them comprising edges to form large-scale networks of science. These network data are the core of much of what I do in my efforts to explore and make sense of the scholarly literature.

The utility of a network view of the literature is intuitive at a small scale: starting from a given publication, one can often find more useful and relevant information by looking at papers that have cited or been cited by that paper. When we zoom out to look at many papers at once, however, this information can easily become overwhelming. This is the case even when the number of papers is relatively small, in terms of modern standards of "big data". Data sets on the order of hundreds of thousands or millions of observations can seem very modest to those who work with petabytes of data at a time. However, in modeling the networked relationships, we quickly start to encounter combinatorial explosion, in which the many interdependent relationships between entities leads to exponential increases in complexity, and what was once merely medium data suddenly becomes huge data. This is where the tools of network science can come into play.

Community detection, or clustering,[8] algorithms can reveal patterns and relationships in complex citation networks. In these networks, vertices represent research papers, and directed edges represent the existence of a citation from one paper to another. The goal of community detection is to identify a way of grouping the vertices such that within-group edges tend to be more prevalent than between-group edges. There are many algorithms available that can be used to detect communities in networks, representing several different approaches to

[8] I use the terms "community detection" and "clustering" interchangeably, and consider "community," "cluster," and "module" to be synonyms. This is the view taken by [67] and [68], and is a view that is fairly well accepted.

the problem. These algorithms are often computationally difficult and with the continually increasing number of publications, the challenge is to adapt these algorithms to very large networks.[9] For context, myself and others in the Datalab have tried unsuccessfully to use traditional, hierarchical community detection methods on large citation networks such as WoS and MAG. Even using fast algorithmic approaches on powerful, high-memory computers, these attempts did not complete after days of processing. I have estimated that it would require weeks to months of constant processing to complete, if they are able to complete at all.

To address these issues, I have developed new methods to cluster very large citation networks. Using several parallel processing techniques, I am able to perform clustering on networks with hundreds of millions of publications and over 1 billion citation links.[10] These are, to my knowledge, the largest networks on which these techniques have been successfully employed. In this section, I describe these methods, and detail some of the ways I have applied them.

2.2.1 Infomap, the Map Equation, and RelaxMap

There are a number of different approaches to the problem of detecting communities in network data. A recent paper by Schaub et al. laid out a taxonomy of four broad categories of these approaches: (i) the *cut-based perspective*, (ii) the *(data) clustering perspective*, (iii) the *stochastic equivalence perspective*, and (iv) the *dynamical perspective* [156]. The different perspectives represent different approaches to the problem, often with different kinds of data, different methods, and different goals. They also represent, to some degree, the different research communities that have been working on the problem. Each of these perspectives have their own set of widely used methods, such as the modularity-optimizing Louvain algorithm for the clustering perspective, and the stochastic block modeling (SBM) techniques for the

[9] An alternative would be to trim the network to an area of interest, but this can introduce sample bias, and a global map of a network can give better results than a local one [99].

[10] Code available at https://github.com/h1-the-swan/infomap_large_network

stochastic equivalence perspective. I focus my work on the dynamical perspective, which is well suited for citation networks as it is primarily concerned with the flow between nodes of the network, an important aspect when modeling information flows through citations. Specifically, I make use of the map equation quality function, and the Infomap algorithm which optimizes it to find a community structure. This algorithm has been shown to be among the best in empirical comparative evaluations of community detection algorithms on synthetic and real-world benchmark networks [103].

The *map equation* uses elements of information theory to describe a random walker's movements on a network. It does this by quantifying the amount of information (in bits) it takes to describe the walker's movements given a particular modular structure of the nodes. The goal is to exploit the relationship between compression and relevant modular structure of a system. The optimal compression means that the smallest number of bits of information is needed to describe the random walker's movements. The map equation is the minimal description length (MDL) of the diffusion process modeled by the random walker given a modular structure.

To make this idea clearer, an analogy can be made to the world of cartography, which is where the map equation derives its name. Assigning an address to a location in a city is a way of encoding that location. Parts of an address can be reused in different locations—there can exist multiple Main Streets in different cities. Reusing these names makes describing an individual's movement within a city more efficient (fewer bits), since as long as we know the individual is staying within the city limits we can just use street names without causing confusion. The map equation formalizes this idea mathematically, providing a function that can be optimized to find good community structure given a network and a model of the dynamics on that network. The method does not actually find the description of the network (the codewords that would be used to compress the network), but rather the lower theoretical limit on the number of bits that would be needed to do the encoding.

Information theory states that this lower bound for the average length of a codeword to describe a random variable X with n possible states that occur with frequencies p_i is equal

to the entropy of that random variable: $H(X) = -\sum_{i=1}^{n} p_i \log p_i$ [50]. (It is standard to use base-2 for the logarithms, which yields calculations in bits.) The map equation imagines that there are separate codebooks for each module (community), and is thus the combined entropy of each codebook plus an additional index codebook that allows for switching between modules, rated by their rates of use:

$$L(\mathsf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathcal{P}^i)$$

where M is the module partitioning; the left term is the average length of codewords (entropy) in the index codebook weighted by the rate of use of the index codebook $q_{\curvearrowright}$; and the right term is the average length of codewords in module codebook i weighted by the rate of use of this module $p^i_{\circlearrowright}$.

The node-visit probabilities that define the p's and q's in the map equation can be calculated using the PageRank algorithm [133], originally developed to rank web pages in the hyperlinked world wide web. PageRank is essentially eigenvector centrality—the node-visit probabilities after an infinite random walk are equal to the leading eigenvector of the adjacency matrix (the eigenvector associated with the largest eigenvalue). PageRank includes an additional element: at any step of the random walk, with a given probability, the walker can "teleport" to a random node on the network. This teleportation is necessary to avoid the possibility of the random walker getting stuck forever in an area of the network with no outward links. The random walker represents an ergodic Markov chain—a time process in which the state at the next step only depends on the current state, and the probability distribution of all states as time approaches infinity is well defined.

The algorithm used to find the community structure that optimizes the map equation is called Infomap, and the code is available at `http://www.mapequation.org/`. It works similarly to the Louvain algorithm used to optimize modularity [20]. Each node starts out in its own module; the modules are joined in a greedy fashion (examining each node one at a time, in a random order) to yield the largest increase of the map equation. These modules

are joined into super-modules, the network is reconstructed with these super-modules as the nodes, and the process is repeated until no further improvement can be made. The method is then extended to find a hierarchical partition [154].

The greedy Infomap algorithm is relatively fast when compared to other methods of finding the optimal value for the map equation, such as considering every possible partitioning. However, since the map equation must be calculated at the level of the full network at each step of the greedy search, the computation cannot be parallelized across multiple processors. For large networks, this problem leads to unreasonably long runtimes. Bae et al. proposed a modification of the Infomap algorithm called RelaxMap, which allows multiple processors to perform the greedy search in parallel.[11] It does this by assuming that a change in module structure near one vertex will not have a meaningful effect on a set of p other vertices chosen at random, provided that p is much less than the number of nodes in the network. This assumption is reasonable if the network is sparse, which is the case for many real-world networks (including citation networks). Using p processors running in parallel, with a lock applied until all processes are done at each iteration, this approach allows for fast, parallel identification of the optimal map equation for a two-level partition of a large graph—meaning a non-hierarchical, non-overlapping clustering of each node into exactly one of any number of clusters.[12]

[11] RelaxMap is a parallel-processor approach to network clustering. Bae and Howe have also proposed a distributed memory implementation called GossipMap [11]. I have not found the need to make use of these methods, as affordable high-memory machines have largely kept up with the size of network. As argued in [137], there are benefits to keeping analysis of large graphs within a single machine and not relying on distributed-memory algorithms. The limitation is not the ability to fit the graph in-memory; rather, the bottleneck is in the computation time of the clustering.

[12] I have done some work, in collaboration with undergraduate students whom I have mentored through the UW Research Computing Club, on maintaining and updating the RelaxMap software. We have also started on extending the functionality of the software to allow for different types of network dynamics, such as undirected links.

2.2.2 Parallel hierarchical clustering

My primary contribution here is a method that combines RelaxMap and Infomap to obtain a hierarchical clustering of a very large network. With RelaxMap alone, we can perform clustering on graphs too large for the Infomap algorithm, but we lose the ability to identify a cluster hierarchy. This is unfortunate, as the cluster hierarchy can often be valuable when looking at networks. For example, the clustering of article citation networks can define high-level scientific topics at the top levels of the hierarchy, and sub-fields within those topics as we go further down the levels. In order to obtain a multi-level clustering of very large networks, I propose a two step approach. The first step uses RelaxMap to identify a two-level (non-hierarchical) clustering of the full network. The second step uses multiple instances of the serial Infomap algorithm, run in parallel on each of the clusters identified in the first step, to obtain a hierarchical clustering of each of the high-level clusters. This assumes that a reasonable (i.e., optimal or close to optimal) hierarchical partition of the full network would have the same clustering at the highest level as the non-hierarchical partition identified by RelaxMap.

I implement the parallel hierarchical clustering step using Apache Spark, an open-source framework for large-scale parallel data processing. The full edge-list of the graph is filtered to only include edges in which both source and target vertices are in the same RelaxMap cluster. (Nodes in small clusters are excluded, as they do not need to be further broken down.) The edge-list is grouped by cluster, resulting in subgraphs for every cluster. Each subgraph has the hierarchical Infomap algorithm run on it to yield a hierarchical partition. Because each of these instances of Infomap runs independently on each subgraph, they can be run in parallel across multiple processors. This process yields, for every node, a hierarchical cluster address. Finally, the original RelaxMap cluster ID is prepended to the cluster address, and the nodes in small clusters are merged back in. The final result is a multi-level clustering of the original network resembling the normal output of Infomap, with every node assigned a hierarchical address of variable depth.

Network	MAG 2019-11	WoS 2019-12	Twitter COVID RT
# Vertices	108,676,190	163,830,918	68,190,226
# Edges	1,554,240,404	1,269,262,278	887,525,809
Runtime Step 1 (RelaxMap)	3.25 hours	4.62 hours	2.54 hours
Runtime Step 2 (Parallel Infomap)	56.27 min	31.07 min	2.11 hours
# Processors	64	64	32
Memory	512 GB	512 GB	187 GB

Table 2.1: Runtimes for getting hierarchical clusterings of large networks, using the two-step parallel method. "MAG" and "WoS" refer to the article citation networks for Microsoft Academic Graph and Web of Science. The Twitter network is a retweet network of users tweeting information about COVID-19. "Memory" refers to the amount of RAM available on the machine used, which may be more than the amount of memory required to process the data set.

2.2.3 Applications, computation, and runtime

Using this method, I have been able to cluster the Web of Science and Microsoft Academic Graph data sets, which previously were too large to cluster. I ran this on the MAG citation network, for example, in about 4 hours on a high-memory compute machine on Amazon Web Services, thanks to support from UW's Research Computing Club. I used a memory-optimized r5.16xlarge EC2 instance with 64 processors and 512 GiB memory, which by current pricing costs less than $25 for the time needed, making this an affordable way to cluster these large networks.

An interesting finding to come out of this work is that it is not always easy to identify whether a given network will be "too big" for Infomap and require the parallel clustering methods. While it is certainly the case that these networks tend to have a large number of vertices and edges, there is no specific number above which we can reliably predict the resources and time needed. Furthermore, there are likely other network properties, such as

the complexity of the community structure, that can affect the computation in various ways. Working with real world data can require some trial and error to successfully employ these methods.

I have leveraged the cluster information I was able to get for these citation networks in several ways in different projects further down the pipeline. I used it as a feature in paper ranking/recommendation models (Autoreview, chapter 3). I used it as well to explore relationships in and between fields in visualizations (such as the intersection of InfoSec and ethics, page 83); and as part of a citation analysis for the Military Suicide Research Consortium (MSRC), identifying different clusters that have been influenced by the research to come out of this group. I have also been able to apply the large-scale clustering method more broadly. For example, I have used it to perform clustering on a Twitter retweet network with hundreds of millions of relationships between tens of millions of users. See Table 2.1 for a summary of using the method on the different data sets.

Overall, my work in maintaining, linking, and processing these data has built up to a comprehensive understanding of how they can best be used, and has formed a solid foundation for other projects. The remainder of my dissertation will cover some of these projects. These include Autoreview, which leverages multiple facets of the data to provide recommendations for papers in a given topic, and several visualization tools I have built on the data to facilitate understanding and exploration.

Chapter 3

AUTOREVIEW

3.1 Author Preface

The literature review, a type of scholarly publication that synthesizes and highlights existing findings that have been previously published, serves a crucial function in scientific progress. By collecting some of the most relevant literature for a given topic, they help interdisciplinary scholars or anyone who wants to find their way in a field outside their expertise. As the size of the literature continues to increase, these reviews become even more valuable. However, this curatorial process does not scale well, requiring a large investment of time and effort. Automated methods have the potential to assist in this process, but a lack of ground truth makes it difficult to develop and evaluate these techniques.

My dissertation's main methods contribution is Autoreview, an approach to this problem that leverages the references in existing review papers as an approximation to ground truth. It rests on the assumption that references in a review represent at least a subset of papers relevant to a given topic. Using this abundant labeled data within the thousands of reviews in the literature, I am able to frame the collection of a literature survey as a supervised learning problem. This allows for experimentation and testing of models and features at a large scale.

The Autoreview work has a peer-reviewed journal article published in the journal *Scientometrics*, with the title, "Constructing and Evaluating Automated Literature Review Systems" [143]. I present the article below as this chapter of my dissertation.[1] In this work, I implemented the framework by training classifiers on 500 review papers, and systematically evaluated and compared different methods, with the goal of selecting the reference papers

[1]This study is previously published work. To cite material from this chapter, please cite the original work: Jason Portenoy and Jevin D. West. "Constructing and evaluating automated literature review systems." In: *Scientometrics* (June 3, 2020). ISSN: 1588-2861. DOI: 10.1007/s11192-020-03490-w

from the review papers out of a large set of candidate papers. By training classifiers using network clustering information combined with title similarity scores, the method was able to achieve an average R-Precision score of 0.385, meaning that the "reconstructed" reference list generated by the model contained on average more than one third of the papers that were in the true reference list. This is a substantial achievement; the task of identifying several hundred target papers from a candidate pool of hundreds of thousands or millions is a difficult challenge for any method, automated or manual. I also showed that the framework allows for development and testing of models and features to incrementally improve the results, and that models I built are able to identify relevant papers even when starting with a very small set of seed papers. In addition to the publication, I make the code freely available.[2]

Beyond these experiments, I have adapted the methods as a tool for generating a list of novel papers relevant to a given field. In this use case, the confidence scores applied by the classifier help find similar papers that were not actually in the target set. In the classic classification task, these would be considered misidentified, but here we consider the possibility that their similarity to the seed papers may make them relevant papers for this field. I have applied this approach to several different fields, and have been able to evaluate the methods by collecting relevance judgments from domain experts. These evaluations came as part of several symbiotic partnerships with outside organizations who have found these methods useful. One of these is with the National Academy of Sciences, with whom I have explored the fields of misinformation studies,[3] science communication, and science literacy; some of these evaluations are included in the Autoreview journal publication. Another is the Military Suicide Research Consortium, to whom I have recommended papers around suicide prevention in the military. They are currently evaluating these recommendations, and I will add their judgments to the overall evaluations of the methods.

Other next steps for the project will include improving the code to make it easier for others to use, and a robustness analysis to explore how stable the automated methods are

[2]https://github.com/h1-the-swan/autoreview

[3]I have created a website to show this work: http://www.misinformationresearch.org/

when identifying new relevant papers. I also intend to work on presenting the new papers in helpful ways using visualizations and other interactive tools; I elaborate more on this in chapter 4.

——————— PUBLISHED WORK BEGINS HERE ———————

3.2 Abstract

Automated literature reviews have the potential to accelerate knowledge synthesis and provide new insights. However, a lack of labeled ground-truth data has made it difficult to develop and evaluate these methods. We propose a framework that uses the reference lists from existing review papers as labeled data, which can then be used to train supervised classifiers, allowing for experimentation and testing of models and features at a large scale. We demonstrate our framework by training classifiers using different combinations of citation- and text-based features on 500 review papers. We use the R-Precision scores for the task of reconstructing the review papers' reference lists as a way to evaluate and compare methods. We also extend our method, generating a novel set of articles relevant to the fields of misinformation studies and science communication. We find that our method can identify many of the most relevant papers for a literature review from a large set of candidate papers, and that our framework allows for development and testing of models and features to incrementally improve the results. The models we build are able to identify relevant papers even when starting with a very small set of seed papers. We also find that the methods can be adapted to identify previously undiscovered articles that may be relevant to a given topic.

3.3 Introduction

Conducting a literature review, or survey, is a critical part of research. As the literature continues to grow and as scholars continue to move across disciplines, synthesizing and highlighting existing findings becomes increasingly important. At the same time, it has become increasingly difficult to identify even a slice of the relevant papers for a given topic [166]. The problem is that this curatorial process does not scale well. It is expensive

in both time and human effort. The advent of Big Scholarly Data—the availability of data around published research and the techniques and resources to process it—has led to a flurry of activity in finding automated ways to help with this problem [4, 16, 93, 152, 197, 211].

Many methods have been developed to recommend relevant papers, using features related to textual similarity, keywords, and structural information such as relatedness in a citation network [16]. However, a common problem in developing and evaluating these methods is a lack of ground truth. We don't know whether our methods are actually selecting relevant papers or topics. This is a general problem in recommender research, but especially so for scholarly papers, given the specialized knowledge needed to evaluate quality and relatedness.

In this paper, we present an approach to this problem that leverages the references in existing review papers as an approximation to ground truth. We assume that references in a review represent at least a subset of papers relevant to a given topic. Using this abundant labeled data within the thousands of reviews in the literature, we are able to frame the collection of a literature survey as a supervised learning problem. Within this supervised framework, we are able to evaluate, at least to some degree, the quality of methods aimed at automatically synthesizing scientific knowledge.

With this framework in place, we demonstrate how supervised learning models can be used to identify relevant papers for review, deriving features from the metadata associated with an article. These features include citations and the groups of papers that can be derived by clustered citation networks [67]. They also include text features derived from the similarity in paper titles. However, any set of related features (authors, disciplines, etc.) could be incorporated.

Using the reference list from a single review article as a benchmark, we develop methods for recapturing those references automatically using the features noted above (Section 3.6.1). We then extend this method beyond one review article and apply the methods to a large group of review articles (Sections 3.6.2 and 3.6.3). Finally, we apply the methods to identify relevant papers in the fields of science communication and misinformation studies. We invite

domain experts to validate our results (Section 3.6.4). We make code and sample data for this project freely available at `https://github.com/h1-the-swan/autoreview`.

The main contribution of this work is a novel framework for constructing and evaluating automated methods for generating references for literature surveys at a large scale. This work builds off of a BIRNDL workshop paper presented at SIGIR 2019 [142]. We have extended this work in several ways: running thousands of experiments to assess how the methods perform using various review articles, sets of features, and data splits; expanding the background literature review; and reporting results from expert feedback on our exploration of new fields.

3.4 Background

There have been several previous attempts at automated or semi-automated literature surveys [18, 44, 89, 90, 159]. These approaches have tended to be smaller scale and rely on more qualitative means of evaluations, which are difficult to replicate and compare across studies. For example, Chen [44] developed a system to aid in writing literature reviews, which was evaluated by having first-year graduate students use it to help them write and submit papers for publication. These student-submitted papers had a high acceptance rate, and one student won a best paper award. This evaluation approach, while creative and compelling, does not scale well. Another study by Silva et al. [159] applied community detection on citation networks to map papers in two different topics and then apply text analytics to generate taxonomies of terms. This approach allowed for detailed analysis of how subtopics are related within a field, but it relied on keyword searches, which can be an insufficient method of identifying all relevant articles [75, 86, 105].

Recent work has explored the use of review articles as a way of testing automated literature review systems. Belter [18] used a semi-automated technique to retrieve documents for systematic reviews using citations. Janssens and Gwinn [89] used co-citation and direct citation networks to identify eligible studies for existing biomedical systematic reviews, starting from one or two known articles. These methods have begun to be used in helping to

create new systematic reviews (e.g., [3]). Other studies have used active learning approaches to reduce the workload associated with selecting relevant articles for systematic reviews in the domains of medicine and public health [117, 180], law [49], and software engineering [205, 206].

Automatically identifying papers for surveys is similar to recommending papers, more generally. This topic has been extensively studied within and outside big scholarly data. A recent survey paper on research paper recommender systems [16] identified more than 200 articles on the topic published since 1998. The survey notes that the majority of approaches use keywords, text snippets, or a single article as input. Our approach, in contrast, starts with a set of seed papers, which are then expanded upon. Our approach also has the distinction of being able to make use of any combination of various features, enabling us to use both textual and network-based features. Some previous work has built recommender systems which combine text and citation information [76, 101]. These take a different approach, using embeddings to characterize similarity between articles.

The new research in automated methods for literature reviews is the result of people applying newly available data and computational power to a perennial and worsening problem—that of the need for and difficulty of organizing large bodies of research. This need for efficient literature review, and especially systematic review, is strongly felt in medicine, but it is also a need for all areas of science [14, 131, 166]. In our work, we aim to provide a framework to help with this research by offering a way to develop and test literature review generation and recommendation at a large scale.

3.5 Data and Methods

3.5.1 Data

The network data used in our analysis come from a recent snapshot of the Web of Science (WoS) citation index consisting of 1,269,262,278 directed citation links between 163,830,918 papers. The data set contains paper-level metadata, such as titles, abstracts, publication

dates and venues, and authors. We used WoS because it is one of the most comprehensive bibliographic datasets, covering a large number of articles across most scientific fields. WoS also identifies certain articles as review papers, which was convenient for this project.

We removed some papers from the full data set. In order to reduce the network to a size that we could cluster (see Section 3.5.3), we removed all papers that had no outgoing citations, and any paper that was only cited once (many of these actually appeared to be placeholder data, for which WoS could not fully identify the cited paper). We also removed papers which were missing all metadata, such as publication year and title. This cleaned data set had 55,271,946 papers, and 1,020,164,414 directed citation links.

3.5.2 Identifying candidate papers and setting up the supervised learning problem

Our procedure is presented in Fig. 3.1. The first step is to randomly split the papers into a set of "seed" papers and a set of "target" papers. We are imagining a researcher who is starting with a set of papers relating to a topic (the seed papers). This researcher wants to expand this set to find the other relevant and important papers in the topic. The target papers can be thought of as the set of papers the researcher has not yet included. Ideally, we would like to search for these target papers within the total set of papers in our data set. However, it is infeasible to generate features and train models using the total set of 55 million papers. To narrow the total set to a more reasonable number of candidate papers, we collect all of the papers that have either cited or been cited by the seed papers. We then go one more degree out, taking all of the papers that have cited or been cited by all of those. We follow a second degree of citations because following direct citations is often not sufficient to identify all relevant literature [89, 151]. This process of following in- and out-citations imitates the recommended practice for a researcher looking for papers to include in a survey, but at a larger scale [188]. The resulting set of papers, while large (generally around 500K to 2M), is manageable enough to work with. We have found that this method, using different samples for the seed papers, reliably generates sets of papers that contain all or nearly all of the target papers (see Section 3.6 and Table 3.2). We label each candidate paper positive

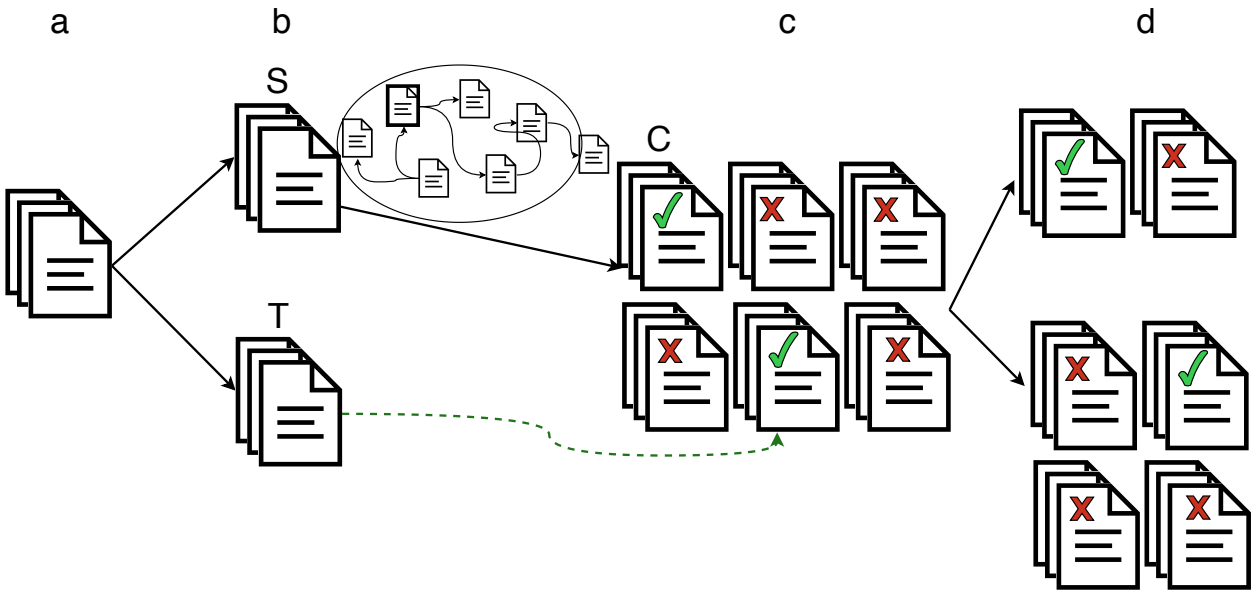

Figure 3.1: Schematic of the framework used to collect data for development and testing of a supervised literature review classifier. (a) Start with an initial set of articles (i.e., the bibliography of an existing review article). (b) Split this set into seed papers (S) and target papers (T). (c) Collect a large set of candidate papers (C) from the seed papers by collecting in- and out-citations, two degrees out. Label these papers as positive or negative based on whether they are among the target papers (T). (d) Split the candidate papers into a training set and a test set to build a supervised classifier, with features based on similarity to the seed papers (S).

or negative depending on whether it is one of the target papers. The goal is to identify the positive (target) papers among the many candidate papers. At this point, we split the candidate papers into training and test sets in order to build classifiers.

3.5.3 Features

Our next step is to generate features to use in a classification model. One feature we use involves the use of citation communities. The structure of the citation network, in which nodes represent papers and edges are directed citations between them, contains patterns about the relatedness of papers. Citation communities are groups of papers that tend to have more connections within-community than outside it. To extract these communities, we

used Infomap to cluster the citation network [153]. Infomap is a well known unsupervised community detection algorithm based on principles of information compression. Identifying clusters in a network of tens of millions of documents is computationally expensive, so we developed a two-step approach to cluster the full network.[4] In the first clustering step, we identified a non-hierarchical clustering of the full network using a parallelized version of Infomap [10]. This process took 5.3 hours on a machine with 32 cores. 5,513,812 clusters were identified in this way. In the second step, we further processed these clusters to identify hierarchical structure, which is something the parallelized version of Infomap cannot do. We wanted to identify this hierarchy because the structure of science tends to be hierarchical, with smaller communities nested within broader ones. To do this, we used Infomap combined with Apache Spark to further cluster all of the top-level clusters with at least 100 nodes into multi-level, non-overlapping clusters. This second step took about 30 minutes on the same machine. The final clustering had 9,390,263 bottom-level clusters, with a maximum depth of 11, and an average depth of 2.9 (std 0.77).[5]

To incorporate the citation clustering information into classification models, we calculate the average cluster distance between a paper and the seed papers. Distance for two papers i and j is defined as

$$\frac{D_i + D_j - 2D_{LCA}}{D_i + D_j} \tag{3.1}$$

where D_i and D_j represent the depth in the clustering tree hierarchy of i and j, and D_{LCA} represents the depth of the lowest common ancestor of the two papers' clusters [54].[6] The

[4]For the clustering, we used the cleaned version of the Web of Science network as described in Section 3.5.1. We used the network after cleaning for citations, but before removing papers with other missing metadata. This version of the network had 73,725,142 nodes and 1,164,650,021 edges.

[5]Since every node is in exactly one cluster (even if the cluster is only one node), and the leaves of the hierarchy tree represent the nodes themselves, the minimum depth in the hierarchy is 2. In this case, the first level is the cluster the node belongs to, and the second level is the node.

[6]We divide the standard measure of distance between nodes in a tree by the sum of the nodes' depth. This is because, in the case of hierarchical Infomap clustering, the total depth varies throughout the tree, and the actual depth of the nodes is arbitrary when describing the distance between the nodes. For example, a pair of nodes in the same bottom-level cluster at a depth of level 5 are no closer together than a pair of nodes in the same bottom-level cluster at level 2.

feature for paper i is the average distance to each of the seed papers. We also use PageRank as a measure of citation-based importance [133].

In addition to these network-based features, we add in a simple text-based feature: the average cosine-similarity of the TF-IDF vector of the paper title to those of the seed paper titles. The purpose of including this feature is to demonstrate how structural- (network) and content- (text) based features can be combined in one model, and can be compared to models with only one or the other. In Section 3.6.3 we extend our analysis to include GloVe word embeddings, and add publication year as an additional feature of paper metadata. There are many other options for features that could be used, including those related to citation or coauthorship patterns, paper text, venue of publication, or any other paper feature that could help identify similarity to the seed papers. Future work will continue this approach, exploring all of these features and how they affect the models' ability to reconstruct the review papers' reference lists. Code and sample data for specifying the features used here are available at `https://github.com/h1-the-swan/autoreview`.

3.6 Results

3.6.1 Application to a single review article

To illustrate how the autoreview process works on a single review article, we use a review article on community detection in graphs [67]. We chose this paper because we are familiar with the topic and could therefore inspect the plausibility of the results. The paper represents a comprehensive review of the topic up to the year of publication (2010). This paper has 262 linked references in our data. We apply the autoreview method using a seed set of 50 papers, randomly sampled from the references. This set of "seed papers" can be thought of as the small set of papers that our imagined researcher above starts with. The remaining 212 papers are "target" papers that we would like to identify.

Table 3.1 shows the results from five splits, each using a different random seed. The "random seed" is an integer that the sampler uses as a starting point; each different random

Table 3.1: Results for autoreview performed on a single review paper, for five different initial random splits of the references into 50 seed papers and 212 target papers. A random forest classifier was trained for each of these splits, for different sets of features. The results shown are for network-based features (average cluster distance and PageRank), and network features + text features (cosine similarity of TF-IDF vectors of paper titles to those of the seed papers).

	Num		Network Features			Network + Text		
Seed	Candidates	Recall	Prec at 10/100/1000	R-Prec	Av Prec	Prec at 10/100/1000	R-Prec	Av Prec
1	633,271	0.986	0.8 / 0.48 / 0.13	0.425	0.27	0.9 / 0.77 / 0.15	0.509	0.472
2	522,098	0.981	0.5 / 0.4 / 0.13	0.415	0.227	1 / 0.72 / 0.14	0.505	0.462
3	828,817	0.981	0.8 / 0.45 / 0.11	0.387	0.23	1 / 0.82 / 0.12	0.5	0.429
4	521,479	0.986	0.7 / 0.42 / 0.12	0.415	0.245	0.9 / 0.76 / 0.13	0.5	0.438
5	1,405,034	0.981	0.8 / 0.47 / 0.11	0.396	0.256	1 / 0.75 / 0.14	0.491	0.474
Avg	782,140	0.983	0.72 / 0.44 / 0.12	0.408	0.245	0.96 / 0.76 / 0.14	0.501	0.455

seed leads to a different split of seed and target sets. Running the process multiple times allowed us to see how the whole system varied when the initial seed papers changed but the review article remained the same. We stopped after five times because generating candidate sets and training models is computationally expensive on the large candidate paper sets. We also wanted to focus our efforts on learning how the system would perform with other review articles (Section 3.6.2).

For each run, we split the 262 papers into a set of 50 seed papers and 212 target papers. After collecting candidate papers, we cleaned the data by removing the seed papers, papers for which we did not have titles, and papers published after the year the review paper was published (2010). Each seed (i.e., each row of Table 3.1) represents one instance of the process in Fig. 3.1. We report the number of candidate papers in the final set for each run. These sets of candidate papers range in size from 500K to 1.4M papers. In each case, only (at most) 212 of these papers are in the positive class. This parallels the experience of a researcher trying to do an effective survey of a topic—the goal is to find the right papers in a large body of literature that can feel overwhelming. With respect to these candidate sets, this method achieves very high recall: 98.3% on average (standard deviation 0.00258).

After identifying seed papers, target papers, and candidate papers, we generated features for each candidate paper, and labeled them according to whether or not they were among the targets. We then split the candidate papers into a training and test set, and trained classifiers to try to identify the targets. We experimented with a variety of classifiers: logistic regression, SVC (support vector classifier), SGD classifier (stochastic gradient descent), gaussian naive bayes, random forest, and AdaBoost classifier [121]. Many of these proved to give poor performance and/or run too slowly, so we proceeded with only random forest, logistic regression, and AdaBoost, selecting the best-performing model for each data set.[7]

Table 3.1 reports evaluation measures for each of these five models, as well as their averages. These include the *Precision at 10, 100,* and *1,000*, the *R-Precision*, and the *Average Precision*. The R-Precision is the fraction of target papers found in the top N papers, where N is the total number of target papers—212 in this case [109]. The Average Precision is the sum of the precision at k for each rank k of a correctly identified target paper, divided by the total number of target papers. Since the fundamental task is to reconstruct the reference list from the original review paper, we focus our analysis mostly on the R-Precision scores, which characterize exactly how well the models were able to do this (see Section 3.6.2 for more discussion on the evaluation measures).

Using two network-based features—the average distance between a paper's cluster and those of the seed papers, and the absolute difference of a paper's PageRank score to that of the average seed paper[8]—a random forest classifier identified, on average, 86 of the target papers (40.8%). We also ran the same experiments using the text-based feature of average paper title TF-IDF similarity to the seed paper titles in addition to the network-based features (see Section 3.5.3). This improved the results: the random forest classifiers then identified, on

[7]Machine learning experiments were conducted using scikit-learn version 0.20.3 running on Python 3.6.9.

[8]Although we only performed ranking and clustering once, it would be ideal to remove all nodes and links past the year of the review paper, as well as the review paper itself, and cluster this network. However, performing a separate clustering for each review paper would be computationally infeasible. Nevertheless, any bias introduced by this should be small, as the clustering method we use considers the overall flow of information across multiple pathways, which makes it robust to the removal of individual nodes and links in large networks.

average, 106 of the target papers (50.1%). In the Appendix, we include some examples of papers ranked by the classifier.

3.6.2 Large-scale study on multiple review papers

We now apply these same methods to multiple review papers. The Web of Science, as part of its citation index data, identifies a subset of articles as review papers (∼1.4 million papers). We run autoreview on a sample of these reviews to explore how the method performs on a variety of review articles, under varying conditions. We first take a random sample of 500 review articles with between 200 and 250 references. We chose this sample as a starting point in order to hold the number of references relatively constant. We believe that review papers of this size represents the type of review for which this method would be useful—a comprehensive review of a well-defined topic. We also explore results on subsets of larger and smaller review articles in Section 3.6.3.

For each of these 500 review articles, we took the references and split them into seed and target sets, and tried to use features of the seeds to predict the targets. Table 3.2 shows summary statistics and results for these papers using the same procedure outlined in the previous section (section 3.6.1). The "number of candidates" refers to the number of papers generated from following two degrees of citations in and out from the seed papers (5 random splits of seed/target for each review paper; 2,500 candidate sets in total). These candidate sets are highly imbalanced, with the 150-200 target papers hidden among hundreds of thousands or millions of candidates; again, this is meant to mimic the predicament of the researcher searching for relevant papers in an large pool of candidates. The candidate sets have very high recall, generally including all or missing just one or two papers (with a few outliers in which the citation-based method failed to capture many of the target papers).

For each review article, we gathered the cited papers, and trained models for 5 different random seeds, corresponding to 5 different splits of the data into seed and target papers. We fixed the size of the seed set at 50 papers (in the following section, we explore the effect of varying the size of this seed set). We chose the best-performing model for each split—in most

Table 3.2: Summary statistics for the 500 review articles, including the number of references per review (i.e., the seed papers + the target papers to predict), the publication year, the number of candidates generated per initial split of the data, the overall recall for the candidate sets, and precision measures for two sets of features—the network features only (cluster distance and PageRank), and network features + TF-IDF similarity of titles.

	min	max	mean	std	median
Number of references	200	249	222	14.4	220
Publication year	1939	2016	2001	15.1	2007
Number of candidates	4476	2,152,834	489,418	348,124	453,386
Recall	0.578	1	0.976	0.0469	0.994
Network features					
—Precision at 10	0	1	0.355	0.248	0.3
—Precision at 100	0	0.69	0.199	0.114	0.18
—Precision at 1000	0.002	0.16	0.062	0.0294	0.057
—Precision at 10000	0.0016	0.0331	0.0109	0.00307	0.011
—R-Precision	0.00625	0.635	0.17	0.0985	0.146
—Average precision	0.000734	0.522	0.0891	0.0735	0.0676
Network + Text features					
—Precision at 10	0.2	1	0.827	0.147	0.9
—Precision at 100	0.06	0.94	0.506	0.129	0.51
—Precision at 1000	0.013	0.173	0.0971	0.0256	0.097
—Precision at 10000	0.004	0.0331	0.0129	0.00248	0.013
—R-Precision	0.0437	0.792	0.385	0.105	0.384
—Average precision	0.00713	0.813	0.306	0.12	0.298

cases, this was a random forest classifier; however, in some instances, a logistic regression or AdaBoost classifier outperformed the random forest.[9]

Again, we report the performance of the classifiers as the *Precision* (at 10, 100, 1,000, and 10,000), the *R-Precision*, and the *Average Precision* [109]. The overall goal is to reconstruct the list of held-out target papers from the reference set of the original review articles. Within this task, the relative rank of the different predictions is not especially important. Because

[9]We chose to report the best-performing model for each experiment, rather than restricting to a single classifier type. This decision did not have a large effect on the results. We chose to be flexible in which classifier to use because there are differences among the different review articles. We will continue to explore the nature of these differences in future work.

of this, we focus the rest of our analysis on the R-Precision scores: the number of correctly predicted target papers among the top N papers, where N is the number of target papers, divided by N. However, looking briefly at the precision at k scores—the ratio of target papers identified at different ranks—we can see that the models do tend to give good performance in terms of ranking relevant papers relatively higher. For example, the models given network and text features had, on average, eight papers correctly predicted among the top ten, and 50 among the top 100. We also report the average precision over all target papers as an alternate measure of precision for all relevant documents. This measure is highly correlated to R-Precision (pearson's $r = 0.97$ across all models), so we focus on R-Precision from this point on for simplicity's sake.

Fig. 3.2 shows the distribution of R-Precision scores for 2,500 classifiers (five classifiers for each of the 500 review articles, each one trained and tested on a different split of the article's references). The figure shows the classifiers that were given both network (cluster and PageRank) and text (TF-IDF-based similarity of titles) features. The average score was 0.385 (standard deviation 0.105); the highest score was 0.792.

Some of the worst performing review articles tended to be year-specific reviews, e.g., "Germanium : Annual survey covering the year 1972". These particular reviews have temporal constraints that the classifiers did not learn well. Publication date was not even among the features available to these classifiers; adding publication year in the set of features available to the classifiers did cause the performance to improve somewhat. However, in future analysis, these year-specific reviews should be excluded if possible, as they represent a less-typical case with a hard constraint on the potential references. Nevertheless, this type of review article only represented a portion of those in the lower tail, so it is only a partial explanation for the poor performance on these papers. The models tended to perform better with smaller candidate sets (pearson's $r = -0.17$ for the relationship between candidate size and score). This is likely due to the fact that these candidate sets simply had less noise by virtue of them being smaller. However, since the candidates are collected based on random splits of the

data, it is not possible to exploit this in order to improve performance (i.e., by limiting the size of the candidate sets).

The analysis to this point has aggregated all review articles together; however, it could be the case that different types of review articles perform differently using these methods. One way to explore this is to look at the discipline of the review articles. Fig. 3.3 shows the same distribution of R-Precision scores as above, broken down by subject. We used the Web of Science subject labels for the review papers (taking the first one if there were multiple), and aggregated them into broad categories. Most of the reviews analyzed were in Medicine (202), Biology (122), and Natural Sciences (101). Most of the subject groups did not perform significantly differently from each other, suggesting that it is no more difficult to predict the reviews' references in, for example, medicine as it is for those in the natural sciences. Some of the groups on either extreme did show statistically significant differences—e.g., engineering did have higher scores than pyschology/social sciences—but in general the differences between groups were modest at most (pairwise independent t-tests, Bonferroni corrected α of 0.0024). It is interesting that we did not find any major differences between fields, given that in bibliometric research, findings often do not generalize across different fields.

3.6.3 Extended analysis

We now extend our analysis to explore how the methods perform under various conditions. The three categories of conditions we experiment with are the number of seed/target papers in the initial split of the review references, the features used by the models, and the number of references in the review papers.

Using the same sample of review papers as in the previous section (Section 3.6.2), we begin by varying the first two of these: the number of seed papers, and the sets of features. We limit our analysis here to a subset of 100 of the previously used 500 articles. This was more computationally tractable, as each combination of seed size and feature set involve training models for five seed/target splits. Fig. 3.4 shows the R-Precision scores for 8 different sets of features and 5 different numbers of seed papers.

Varying features: Each line in Fig. 3.4 represents the performance of classifiers using different sets of features to rank and identify target papers, with better performing feature sets on top. Using only the TF-IDF information for paper titles gave the worst performance (~ 0.1). Using network features alone—either cluster information, or cluster information combined with the paper's PageRank scores—resulted in somewhat higher scores than TF-IDF features (~ 0.15, a 50% improvement over TF-IDF). Combining network and text features, as we saw in the previous section, gave a large boost in performance, with scores around 0.4. Adding another feature from the paper metadata—the publication year—gave another boost, with scores around 0.6.[10] We believe that this improvement is because topics in science tend to be situated in a given period in time. By giving the model information about the publication years of papers, it is better able to identify the important papers in the field.

In order to test more sophisticated text features, we also explored models using title embeddings. For each paper title, we found the average word vector from 300-dimensional GloVe embeddings.[11] We used as a feature the cosine similarity between this vector and the mean of the title vectors for the seed papers. These features tended to perform very well; in fact, using embeddings alone absent any other features tended to give scores higher than most other sets of features that did not include embeddings. The best performing models we tested were ones that combined all types of features—word embeddings, network features, and publication year. These models had R-Precision scores around 0.81.[12]

Varying seed size: Each point along the x-axis of Fig. 3.4 represents results from starting with different sizes of seed/target splits. For example, for each leftmost point, the autoreview process began for each of the 100 review papers by randomly splitting the 200-250 references into a seed set of 15 seed papers and 185-235 target papers, with the target papers

[10]The actual feature used was the absolute difference between a paper's publication year and the mean publication year of the seed papers.

[11]We used the spaCy library (version 2.2.3) with a pretrained English language model (`core_web_lg` version 2.2.5).

[12]The models that had both network and title embedding features, but not publication year ("Cluster, PageRank, Embeddings"), performed worse in general than models with embeddings alone, with scores tending to be between 0.5 and 0.7. The reason for this is unclear.

then used to generate the candidate sets. Again, this procedure was done with five different random seeds for each review, for each seed size (15, 25, 50, 100, and 150).[13]

Intuitively, we might expect performance to increase along with the size of the seed set, since with more seeds, the classifiers have more knowledge of how similar papers should look, and fewer target papers to predict. We do see this trend for some of the feature sets—for example, with network + TF-IDF, and network + TF-IDF + publication year. Notably, for each of these, the scores for the smaller seed sets are only modestly lower than the largest seed sets, which suggests that this method can perform fairly well even with only a handful of seed papers. On the other hand, some of the feature sets do not improve with more seed papers. The classifiers using title embeddings alone is the most extreme of these: these models actually perform best with the fewest number of seed papers, and performance decreases as the number of seed papers increases. While the reason for this is not entirely clear, it may due to a tradeoff between having more seed papers—which means more information for the classifier to use—but fewer target papers—which means the classifier has to identify the target papers higher up in the rankings in order to get a high score.

Varying size of review papers Fig. 3.5 shows the average R-Precision scores when starting with review articles with reference papers of varying length. The medium size articles are the same 100 as above, with a seed size of 50. The small review articles are a different set of reviews that have an average of 50 references, with 15 of these references used as the seed papers. The large reviews are another set of reviews that have on average 945 references, with a seed size of 50. These results are largely consistent with those above. Models with only network or only TF-IDF features all perform about the same, regardless of review paper size. For other feature sets, small review papers tend to perform better than larger ones, but this may be a function of the ratio of seed papers to target papers (as seen in many of the models in Fig. 3.4), and not due to any inherent differences between these groups of review papers.

[13]Since the same random seeds (1, 2, 3, 4, 5) were used each time, the smaller seed sets are always subsets of the larger ones. For example, for a given review article and a given random seed, the 100 seed papers identified are all included in the set of 150; the set of 50 seed papers are all included in both the set of 100 and 150; and so on.

3.6.4 Exploring scientific fields using automated literature review

The method we introduce can be adapted as a tool for exploring key papers in an emerging field. In this use case, it is the papers the classifier "misses" that we are interested in. The classifier, attempting to predict the target papers, assigns a confidence score to each of the candidate papers. We are interested in those candidate papers which received a high score, yet were not actually target papers. In the classic classification task, these would be considered misidentified, but in this task we consider the possibility that their similarity to the seed papers may make them relevant papers for this field. This is consistent with Belter's suggestion of "supplement[ing] the traditional method by identifying relevant publications not retrieved through traditional search techniques" [18]. As a case study, we applied this method to papers in the emerging field of misinformation studies, which pulls research from psychology, risk assessment, science communication, computer science, and others.

As part of this case study and in collaboration with the National Academy of Sciences (NAS), we curated a collection of important papers in this field[14] and used this collection as a seed set to identify other related papers that might have been missed by our more manual methods. Evaluating these results brings us back to shaky territory where we do not have ground truth. However, conversations with domain experts interested in formally characterizing these fields have been encouraging, suggesting the utility of these methods in identifying relevant papers. The original seed papers and the extended bibliography of machine-identified and ranked papers can be found at `http://www.misinformationresearch.org`.

Leveraging the expertise of the NAS scientists, we are also studying how well these methods can identify papers in a somewhat more established field. We used a seed set of curated papers in the field of Science Communication to identify and rank additional papers. The seed set consisted of 274 papers collected from a 2017 National Academies report on science communication [125]. We performed five different splits of these papers into seed and target sets (see Fig. 3.1). For each of these, we generated large sets of candidate papers from Web of

[14]See Data and Methods at `http://www.misinformationresearch.org` for details

Science, and then trained random forest models to rank candidates based on the citation- and title-based features described in Section 3.6.1. For each candidate paper, we aggregated the results of the five classifiers by taking the sum of the models' predicted probabilities. We then provided the evaluators with a list of the top 1,000 papers for evaluation that were not in the original seed set. Three domain experts have evaluated this data set, one independently, the other two working together. They made binary relevance judgments for each of the 1,000 papers, with the instructions: "identify any references that the algorithm picked up that don't belong in the field of science communication." The first rater judged 947 (95%) of the references to be relevant, while the other two judged 872 (87%) to be relevant (moderate inter-rater reliability between the two ratings: Cohen's $\kappa = 0.37$). We plan to make use of expert evaluations to assess how useful this approach could be in other fields, including misinformation studies.

3.7 Discussion

Our results suggest that it is possible to use automated methods to identify many of the most relevant papers for a literature review, starting from a large set of candidate papers. We believe that, by trying new features and tuning model parameters, we can increase performance and learn more about what distinguishes these papers. We have also seen promise in using these methods to build novel surveys of topics from a set of seed papers. An important area of future work will be collecting more expert-labeled evaluations to validate and improve this approach.

Running these experiments on our samples of review articles required thousands of hours of computation on a supercomputing cluster. However, applying the methods to a single set of references (as in Section 3.6.4), is much less intensive, and does not necessarily require these resources that may not be broadly accessible or scalable to a general audience.

Previous work on automated methods for literature review have tended to use a small number of hand-selected systematic review articles [18], or a small number of scientific fields [159]. The small scale and close qualitative approach can provide a lot of insight, but

makes it hard to specify benchmarks to generalize and compare different methods. Our experimental approach, on the other hand, gathers many review papers and applies general techniques, allowing for a much larger pool of labeled data.

We found that we were able to identify many of the references of review articles in a variety of research areas. Our methods also missed many references, ranking other articles more highly than the ones in the original reference list. However, it seems that these "incorrect" articles may actually have value: they may be relevant articles that were missed by the review papers' authors. We found some support for this with the help of domain experts, who found that many of the "misclassified" articles were in fact relevant to the given field. While the precision scores attainable by these methods represent a good goal when making improvements, it is worth noting that in many cases, the failures of the classifiers may actually indicate valuable papers that have been overlooked.

Furthermore, we see potential in using this framework to develop and evaluate methods for literature survey generation and related problems such as scholarly recommendation and field identification. The objective we propose for our modeling task—accurately finding all of the remaining references from a review paper given a held out sample of seed papers from those references—is not a perfect one. We assume that the references in a review paper represent domain experts' best attempt to collect the relevant literature in a single research topic; however, there exist several different types of review article (systematic review, meta-analysis, broad literature survey, etc.), and our current method ignores potential nuance between them. Additionally, we assume that every article in a review paper's bibliography is a relevant article to be included in a field's survey; in practice, an article can be cited for many different reasons, even within a review article. Despite these limitations, the large amount of available data allows our framework to provide a means of experimenting with and developing methods for automated literature surveys. There are many review articles similar to the ones we used that have their bibliographies available and so it will be possible to do this development and analysis on a large scale across many domains. Using this framework, it will be possible to

empirically evaluate novel features for their use in identifying papers relevant to a survey in a given topic.

3.8 Acknowledgements

We thank Dr. Chirag Shah for helpful conversations around evaluation measures, and Clarivate Analytics for the use of the Web of Science data. We also thank three anonymous reviewers for constructive feedback.

This work was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system and funded by the STF at the University of Washington.

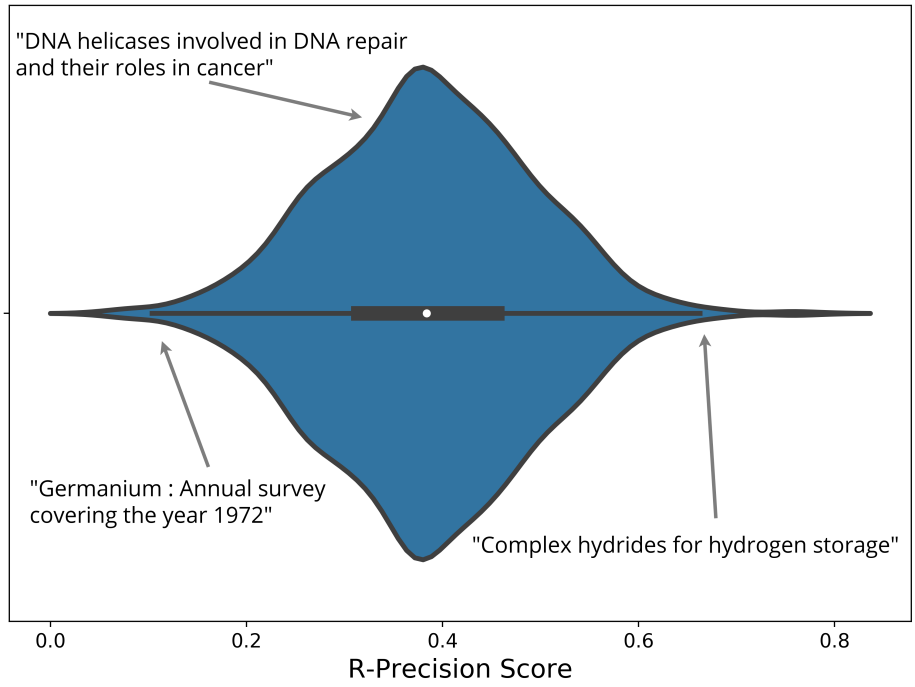

Figure 3.2: Violin plot showing the distribution of *R-Precision* scores (number of correctly predicted target papers divided by total number of target papers) for 2,500 classifiers, each trained on one of 500 different review articles. The violin plot shows a box plot in the center, surrounded by a mirrored probability distribution for the scores. The distribution is annotated with the titles of three review articles. The review article in the lower tail was one of those which the classifiers did most poorly at predicting references (mean score: 0.14). The one in the upper tail is an example of a review paper whose classifiers performed best (0.65). The one in the middle at the fattest part of the distribution is more or less typical for the review articles in our set (0.39).

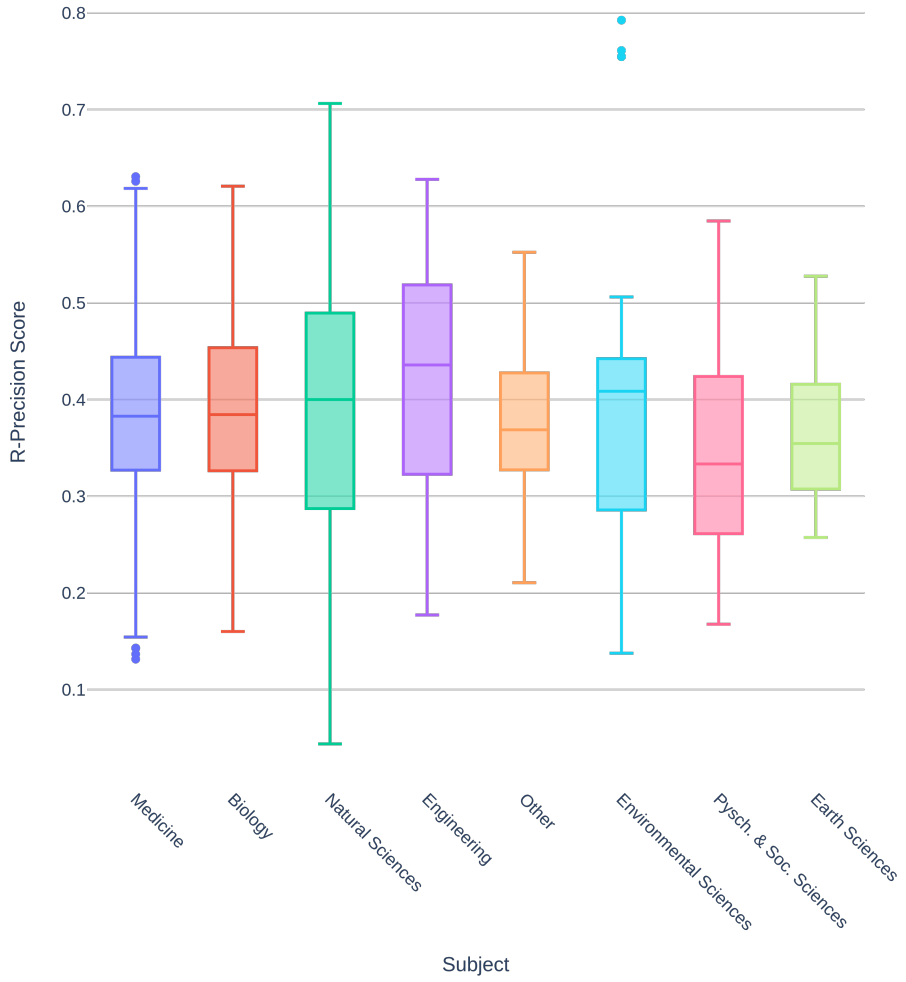

Figure 3.3: Box plots of the *R-Precision* scores for the 500 review articles by subject. 50 seed papers, network and TF-IDF title features. See text for discussion.

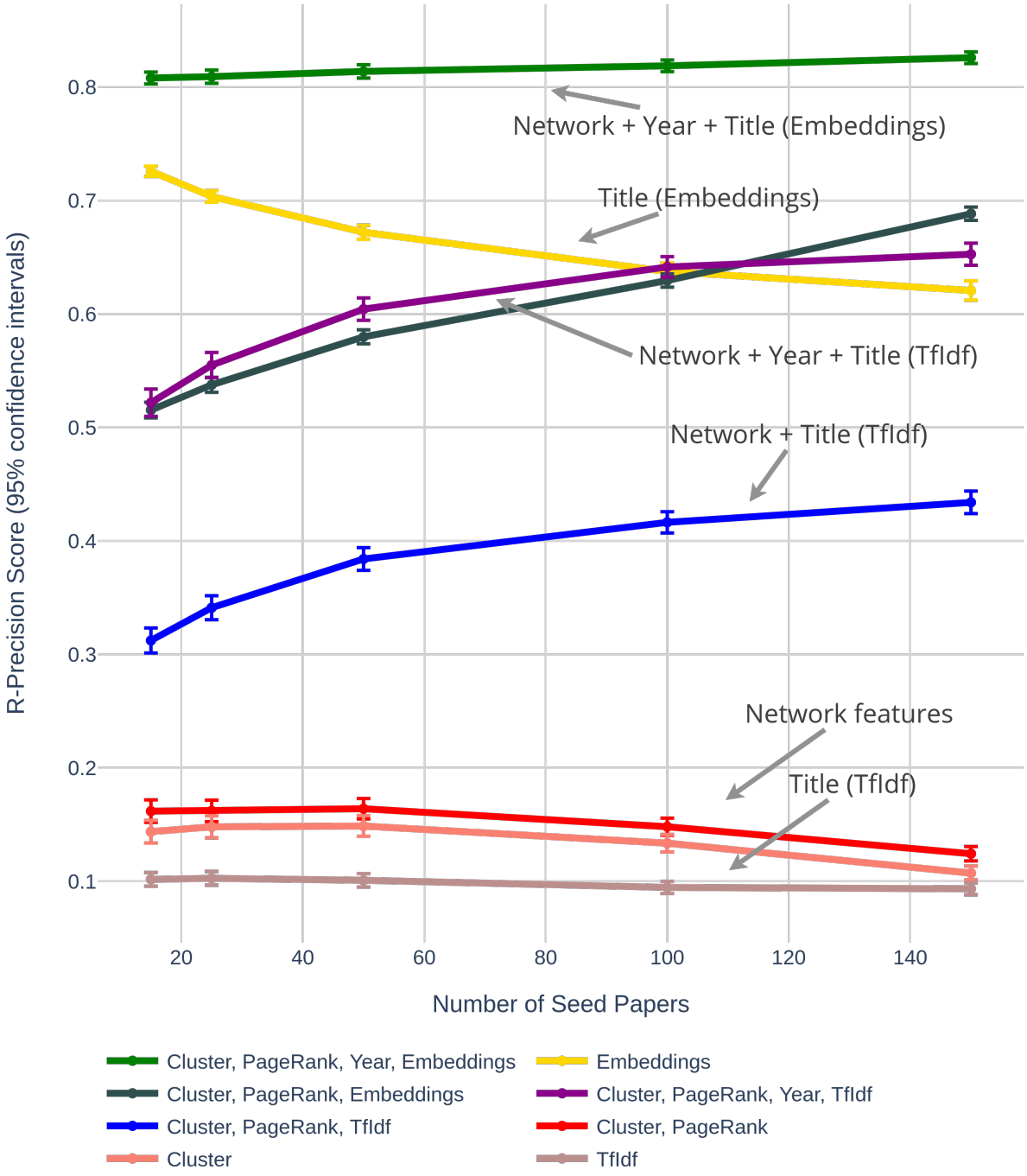

Figure 3.4: R-precision scores for autoreview, varying the number of seed/target papers, and the sets of features used. Each point represents the mean of the R-Precision scores for 500 models—5 each for different seed/target splits of the references of 100 review papers. The error bars represent 95% confidence intervals.

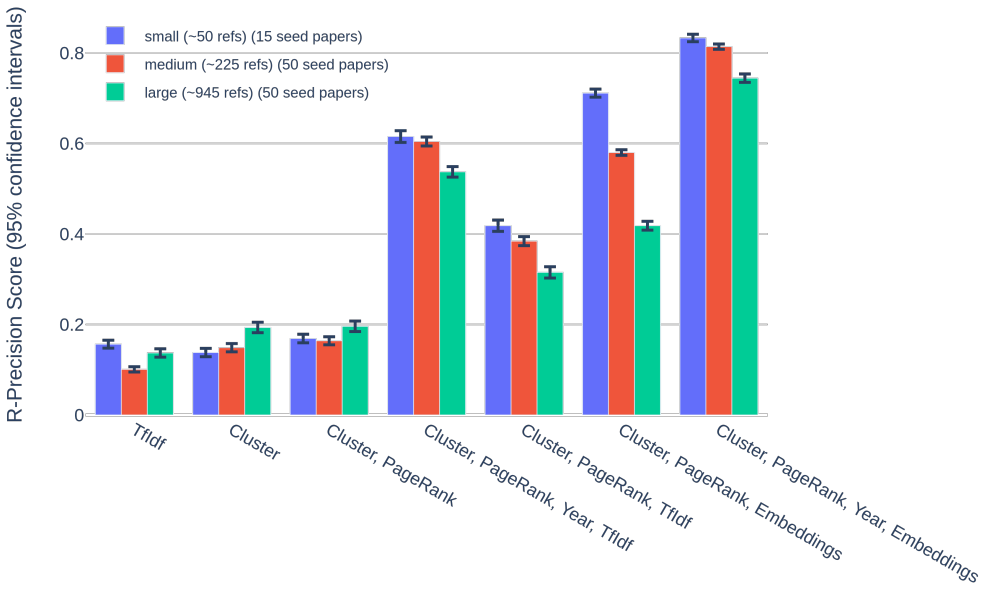

Figure 3.5: Average R-Precision scores for different size review articles. The middle (red) bar for each feature set represents the average score for the same 100 review articles using the same procedure as in Fig. 3.4 (seed size 50). The other two bars in each group represent a different set of review articles, the left a set of 100 smaller reviews (50 references on average), the right a set of 100 larger reviews (945 references on average). Error bars represent 95% confidence intervals.

Chapter 4

VISUAL EXPLORATION AND EVALUATION OF THE SCHOLARLY LITERATURE

Exploring and conveying patterns in large-scale networks is an ongoing challenge, but visualizations can make these patterns clearer [104]. A major focus of my work has been in leveraging the citation connectivity of the literature to develop interactive visualizations of science meant to aid in understanding and exploring the literature. In this section, I present my contribution of several tools I have created with this goal in mind.

There is a large body of research into mapping science [22, 42, 47]. These visual representations have the potential to assist policy makers, researchers, and other practitioners in navigating the literature at a high level [13, 59]. [24] lays out some of these potential benefits for a variety of users, such as helping students understand new knowledge domains, facilitating researchers' search for relevant research or potential collaborations [28], serving as tools for grant agencies and R&D managers to identify new innovations, and improving the communication of science to the general public. However, there exist numerous challenges to doing this effectively, some of them having to do with the *communication* to the users of these maps, and others with the details of *implementation*. With regards to communication—there is a learning curve for most of these tools, and the potential benefits are not always immediately apparent. In terms of implementation—many of the currently available tools are clunky, and the design decisions around what to show, what to obscure, and how to interact are not always ideal [13, 35, 59]. Visualizing groups and community structure in networks such as these—one of the approaches with the most potential, in my opinion—presents its own set of challenges, including issues around scalability and data complexity, time-varying factors, and evaluation [173]. Due in part to these challenges, actual adoption of tools among

practitioners has been low. However, I believe that, as the need to deal with information overload increases, and as the data and methods improve, they could become valuable tools for many researchers, as well as policy makers and science of science researchers. This is why I have focused some of my efforts on this development and design work.

This chapter discusses the interactive visualizations and tools I have created to help explore *collections* of papers—a common thread throughout my research. These collections can represent different concepts: fields of study identified either manually or by Autoreview, authors, or research groups. My first interactive visualization I discuss—the nautilus diagram—explores and evaluates the influence that an author or field has had over time and across fields. Subsequently, I developed several other interactive visualizations to help explore collections of papers: a cluster comparison network diagram, a coauthorship network diagram, an article timeline, and an article citation visualization. I have developed all of these applications as part of collaborations with several organizations, and this has allowed me to design tools to view and explore the scholarly data with specific, interested users in mind. In most of these cases, I have had the opportunity to develop or demonstrate the tools with these users, and have received valuable feedback. Taken together, they serve as a suite of exploratory tools to gain insight into these collections. They also can be viewed as a progression in my personal research path, leading up to my current ongoing work on SciSight. SciSight is the final project I will discuss—it is a visualization tool to explore groups of researchers and their similarities and to expose potential gaps in communication and knowledge transfer.

4.1 Nautilus Diagram: Visualizing Academic Influence Over Time and Across Fields

4.1.1 Author Preface

Assessing the influence of a scholar's work is an important task for funding organizations, academic departments, and researchers. Common methods, such as citation counts or h-index, can ignore much of the nuance and multidimensionality of scholarly influence. In an attempt to provide a richer, more nuanced, and more enjoyable view of impact and influence, I created

the nautilus diagram (Figure 4.1). The nautilus diagram is a visualization tool to represent scholarly influence over time using a network representation of citations between publications. I built this visualization as part of a design study with the cooperation of the Pew Biomedical Scholars program. This prestigious program provides 4 years of early-career funding to approximately 30 researchers in health-related fields each year. They have funded multiple highly influential researchers in the biomedical sciences, including several Nobel Prize winners. The program was celebrating its 30-year anniversary, and its administrators had approached the Datalab wanting to reflect on its history using more than standard bibliographic metrics alone. I conceived the nautilus diagram as a way to provide a richer exploration of the impact and influence of the award winners.

The focal center of the nautilus represents a collection of publications—for example, the papers authored by a particular scholar. The visualization shows important papers that have cited these center papers, along with citation links to the center and to other papers. These papers appear as animated circular nodes, positioned in a spiral (nautilus) with the earliest papers near the center and more recent papers near the periphery. The use of spatial placement to represent time is a unique aspect of this visualization; temporal features are typically not displayed clearly in network diagrams, but this strategy conveys this information in a compact and visually compelling way. Nodes are colored by field of study, which can show the diversity of the focal scholar's influence by the variation of color in the overall presentation. The animated visualization tracks along with more traditional accompanying line graphs which show statistics about the central papers per year—number of papers, number of citations, and citation-based influence (as measured by Eigenfactor score).

Using this visualization, I was able to plot these influence graphs for several hundred of the current and former recipients of the Pew award, ranging from early-career to established researchers. I had the opportunity to demonstrate it to 26 of the researchers in person at the program's 30th anniversary reunion in 2015, watching them view the progress and influence of their career. These user interviews served as an evaluation of the design of the visualization, and I collected valuable feedback as well as interesting stories about the

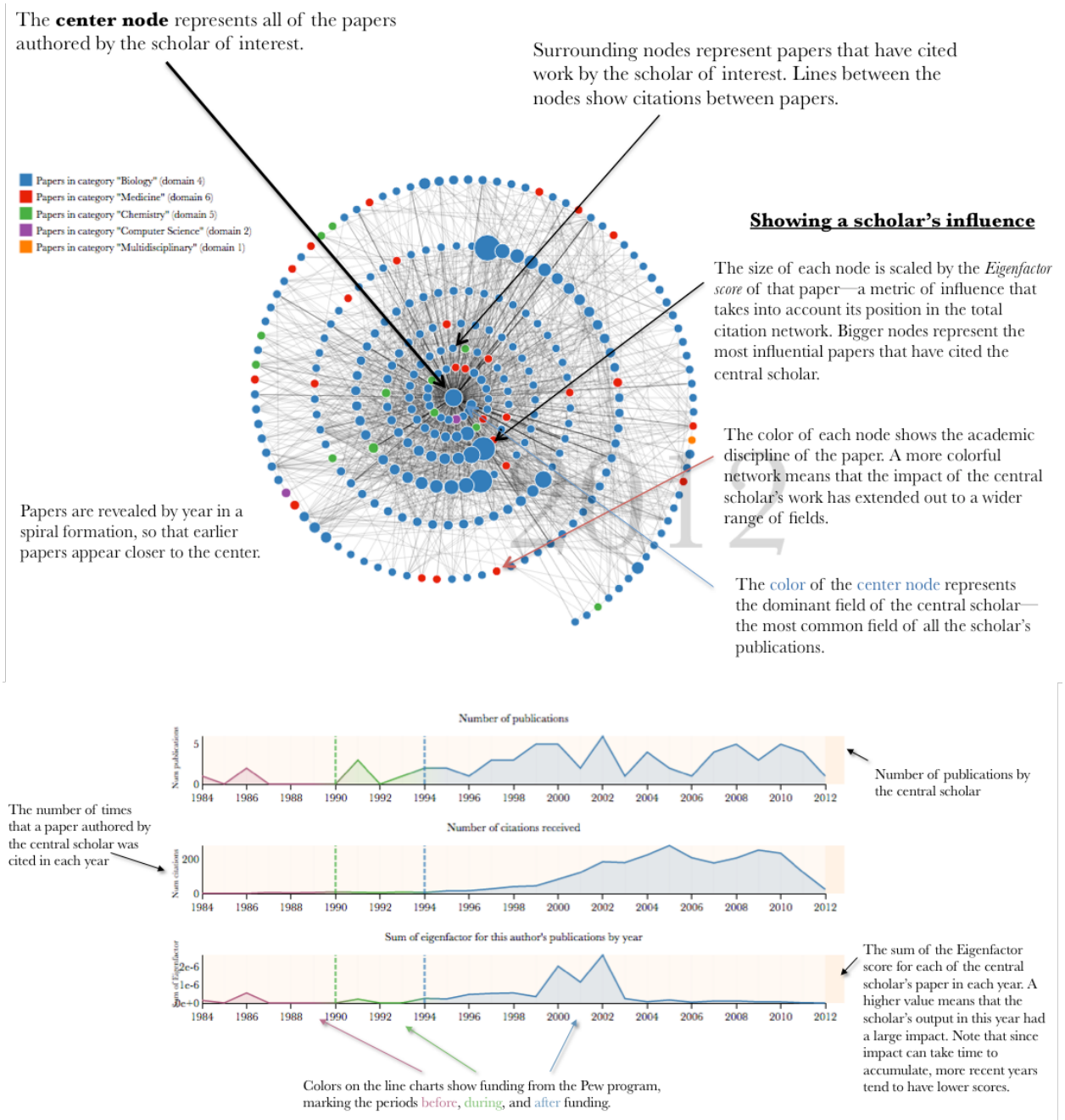

Figure 4.1: Example nautilus diagram, showing the influence of an author over time.

scholars' careers that related to the data being visualized in front of them. I published a peer-reviewed journal article with Jevin West and Jessica Hullman detailing the visualization and its design process and evaluation [140]. This article is included below as the rest of section 4.1.[1] Two subsequent workshop papers also built on this work [144, 145], as well as the web application http://scholar.eigenfactor.org/ . I have also adapted it to other projects, such as visualizing researchers and institutions in the Health Research Alliance, and projects with HICSS and the National Academy of Sciences (described in section 4.2).

[1]This study is previously published work. To cite material from section 4.1 of this chapter, please cite the original work: Jason Portenoy, Jessica Hullman, and Jevin D. West. "Leveraging Citation Networks to Visualize Scholarly Influence Over Time." In: *Frontiers in Research Metrics and Analytics* 2 (2017). ISSN: 2504-0537. DOI: 10.3389/frma.2017.00008

PUBLISHED WORK BEGINS HERE

4.1.2 Abstract

Assessing the influence of a scholar's work is an important task for funding organizations, academic departments, and researchers. Common methods, such as measures of citation counts, can ignore much of the nuance and multidimensionality of scholarly influence. We present an approach for generating dynamic visualizations of scholars' careers. This approach uses an animated node-link diagram showing the citation network accumulated around the researcher over the course of the career in concert with key indicators, highlighting influence both within and across fields. We developed our design in collaboration with one funding organization—the Pew Biomedical Scholars program—but the methods are generalizable to visualizations of scholarly influence. We applied the design method to the Microsoft Academic Graph, which includes more than 120 million publications. We validate our abstractions throughout the process through collaboration with the Pew Biomedical Scholars program officers and summative evaluations with their scholars.

4.1.3 Introduction

The scholarly literature forms a vast network that is connected through citations and footnotes. This well-preserved system—through its billions of links—connects papers, authors, ideas and disciplines over centuries. The structure of this system can reveal where ideas have come from and where they might be going. Though De Solla Price first recognized the potential of citation networks for improving search, evaluation and discovery more than 50 years ago [161], realizing the potential of citation networks for conveying patterns in scholarship has been challenging. Recent advances in data access and scaling pave the way for increased focus on how to communicate the insight captured in citation networks.

One common scenario that calls for ways to accurately and efficiently convey citation network data is measuring scholarly influence. Funding agencies, hiring and promotion committees, and university leaders want to measure the impact of their scholars, but few

tools sufficiently address this task. There have been many proposed metrics for measuring influence [81, 181], but none suffice in capturing the full complexity of scholarly influence. For these aspects, it can be more effective to visualize the movement of ideas between papers via direct citations. There have been many attempts at mapping the scholarly literature using citation networks [47, 202], however, most of these attempts view science at the aggregate, disciplinary level. For this paper, we focus at the local view—at the level of an individual author—with an interest in depicting the *influence* of this author over time. Specifically, we are interested in temporal, author-level citation networks in which the nodes represent papers that cite the work of a particular scholar.

A number of different parties have an interest in looking at the influence of scholarly work and individual scholars. Funding organizations—including nonprofits and government agencies such as the National Institute of Health—collectively spend billions of dollars annually to fund research. These organizations are continually faced with the question of how best to evaluate the impact that the funding has had. University departments tasked with hiring and promotion decisions must evaluate the impact of research as well. Many scholars are interested in looking at their own influence as a means of self reflection, or at other scholars in their field.

The primary contribution of this paper is a broadly accessible, automated, data-driven approach to visualizing the influence of a scholar over time. We apply the approach to the Microsoft Academic Graph, a large (publicly available) citation network. We report on the development of this method through a design study with the Pew Biomedical Scholars program. We validate the design abstractions through demonstrations and discussions with the Pew program officers. We also report on the insights from a validation study in which the Pew scholars themselves interact with the visualization. We extend these methods to offer a publicly available service to visualize scholars' influence at `http://scholar.eigenfactor.org`. We conclude with a discussion of insights gained from this study and future opportunities for work in this space. Our work adds to the growing literature on Knowledge Mapping [123, 41].

4.1.4 Background

Assessing Scholarly Influence

Communicating scholarship at the individual level for assessment has taken qualitative and quantitative forms. More qualitative methods include research narratives authored by a scholar herself, text articles written about a scholar, interviews, or the career retrospectives that occur at conferences and other scholarly events as a way of acknowledging the importance of scholars' contributions. These forms convey a scholar's career in detail in an accessible narrative form. However, these types of reviews take considerable time to prepare and do not easily scale.

Quantitative methods of capturing scholarly impact, often for evaluation purposes, have been used for many years as well. These include measures such as counts of publications and citations. The h-index was introduced in 2005—a researcher's h is the maximum number h so that h papers have each been cited at least h times [81]. Although this measure has received increased attention in recent years as a means of assessment, it still suffers from many problems of its predecessors, such as bias along academic field, academic age, and gender [94, 106].

Another problem with methods that use straight citation counts is they do not take into account the quality of citations. Several methods have been proposed that use the structure of citation networks to algorithmically weight links according to their overall influence (a method analogous to Google's PageRank for websites [133]); these include Eigenfactor [191], Y-factor [21], CiteRank [179], and P-Rank [203]. Our approach employs the article-level Eigenfactor metric, which ranks individual papers according to their position in the network [189]. However, while these methods are considered to be an improvement over simple citation counts, in isolation they can still fall victim to similar issues and biases.

We suggest that using visualization to convey scholarly impact can capture a scholar's influence in a way that provides both qualitative and quantitative information. Visualizations are often used as a means to engage novice and more expert users alike. Visualizations can

make patterns and relationships in a data set clearer [104], and act as storytelling devices [158]. A well-designed visualization can also support analysis to varying degrees of detail, from providing a gestalt view of the overall pattern of a scholar's career while still allowing for more careful examination of the subtler differences in the type or magnitude of influence the scholar has had.

Citation network visualization

There is a large body of work on the topic of mapping and visualizing networks of scholarly publications [47, 41]. Many of the existing techniques define their links using similarity measures—bibliographic coupling, co-authorship, and co-citations. For example, the CiteSpace tool uses co-citation networks and other methods to support visual analytic tasks of science mapping, in order to explore large scale trends in science [40]. Relatively less work has been done visualizing direct paper-to-paper citation networks. Since we want to look at the *influence* of scholars, we are more interested in these direct citation networks than measures of similarity.

There are several tools that do support visualizing direct citation networks, including Action Science Explorer [57], the Network Workbench Tool [23], CitNetExplorer [58], Citeology [111], and PivotPaths [56]. While some of these tools offer the ability to view a particular paper, including author selection, they are designed to support analysis of a network at a particular point in time. Our approach, in contrast, views a dynamic network over time to tell a story of changing and developing influence over the course of a career.

Visualization of dynamic networks

Dynamic network depiction is a particularly challenging subset of network visualization due to the need to show changing structure while preserving the mental model of the user. Animation naturally affords interpretation of change over time [169]. However, to ease the cognitive cost of maintaining the mental model requires limiting change to node positions over time steps and/or smoothly interpolating node positions between frames [119, 149]. Our

technique avoids the difficulty caused by changing node positions by using a radial layout with a fixed anchor point[119] from which new nodes (representing chronologically published papers by a scholar) spiral outward, encoding time redundantly with the animation.

Radial layouts have been used as a way to retain context by snapping nodes of interest to a central point to facilitate analysis centered on different nodes [204]. Applications that map time to the distance from the center point are less common, though several static layouts are exceptions. TimeRadarTrees [31] encodes changes across a sequence of graphs as circle sectors. Each circle sector extending outward from a center point represents a subsequent time step, and each sector is divided into as many sections as needed to depict the nodes and their incoming edges. TimeSpiderTrees also produce static visualizations of dynamic graphs using radial layouts, but by using orientation rather than connectedness to express relationships between nodes. The result is a sparser visualization in which half links between nodes represent changes [32]. Farrugia et al. use a radial layout in which concentric circles represent time periods in dynamic ego-networks [64]. Radial layouts have also been used to depict an adjacency matrix at multiple times steps as rings of a circle [174]. We similarly use a metaphor of time as distance from the center of a circle to depict network data. However we use a spiral shape based on their ability to act as a metaphor for temporal change [2].

4.1.5 Methods

Context

We began exploring methods for visualizing scholarly influence after being contacted in early 2015 by the Pew Scholars Program in the Biomedical Sciences. This program provides four years of early-career funding to approximately 30 researchers in health-related fields each year. They have funded multiple highly influential researchers in the biomedical sciences including several Nobel Prize winners. The program was celebrating its 30-year anniversary and wanted to reflect on its history using more than standard metrics alone (e.g., citation

counts, h-indexes, impact factors). We met with the program directors to discuss richer ways of exploring their impact and influence on biomedicine.

The Pew Charitable Trust is one of many foundations and funding agencies trying to figure out how to measure their impact on scholarship. We viewed this evaluation as a case study in how to better visualize scholarly influence for individual scholars in general. The Pew scholars have been publishing for several decades, their publication data is readily available in open repositories like PubMed Central, and they tend to be influential scholars from a diverse set of disciplines. This prompted us to consider ways to convey not only how much influence the scholars have had, but also the qualitatively different kinds of influence that a scholar could have.

Based on the Pew program's goal of reflecting on their history and our own perception of a broader opportunity to use visualization to convey scholarly impact, we approached the design study as a case study in using data visualization as a storytelling device. Throughout the design study, we referred to the data on a scholar as a story comprising multiple events. This emphasis on storytelling helped encourage us to explore ways of presenting the data that could make it accessible to users who are not necessarily accustomed to using interactive visualizations for analysis, in the same way that narrative visualizations are used to make data more accessible to audiences in the media and other outlets.

Design Study Methodology

We developed the visualization by using an iterative design process over the course of about five months. We met remotely with the Pew program officers eight times throughout this period. Initial meetings consisted of discussing how to frame the goals of conveying scholarly impact and to brainstorm specific measures and visual presentation styles (e.g., animation, static snapshots). Subsequent meetings were used to demonstrate and receive feedback on the design iterations. This process culminated in demonstrations and testing with the Pew scholars at the reunion conference; this is discussed below in Results.

We followed Munzner's nested model for visualization design and validation [120]. This model characterizes visualization design and evaluation at four nested levels—problem characterization, data/operation abstraction, encoding/interaction, and algorithm—and identifies threats to validity at each level. In the next section, we address the last three levels, describing our design process and addressing threats to validity through justification or evaluation.

4.1.6 Design

Data abstraction

Data set: Our database of scholarly publications comes from a public release of Microsoft Academic Search. The data set for our study contains about 49 million papers and 260 million citations. Papers have associated metadata such as title, year, list of authors, journal or conference, abstract, etc. There is also an assigned domain for each paper (e.g. "Biology," "Chemistry," "Computer Science")—this domain has been assigned by Microsoft at the time of collection.

Since the initial design study, Microsoft Academic Search has been decommissioned and replaced with Microsoft Academic, a service powered by a database called Microsoft Academic Graph (MAG), which indexes scholarly articles using content that search engine Bing crawls from the web [160, 80]. Since this update, Microsoft Academic has been gaining traction as a bibliometric research tool, as it approaches Google Scholar in terms of coverage while having much better structure and functionality for researchers akin to Scopus and Web of Science [87]. Our current design uses a snapshot of this graph from February, 2016 [2]; this updated data set has about 127 million papers and 528 million citations. These new data do not provide a single domain for each paper, but rather an array based on a Microsoft-determined

[2] downloaded from `https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/`

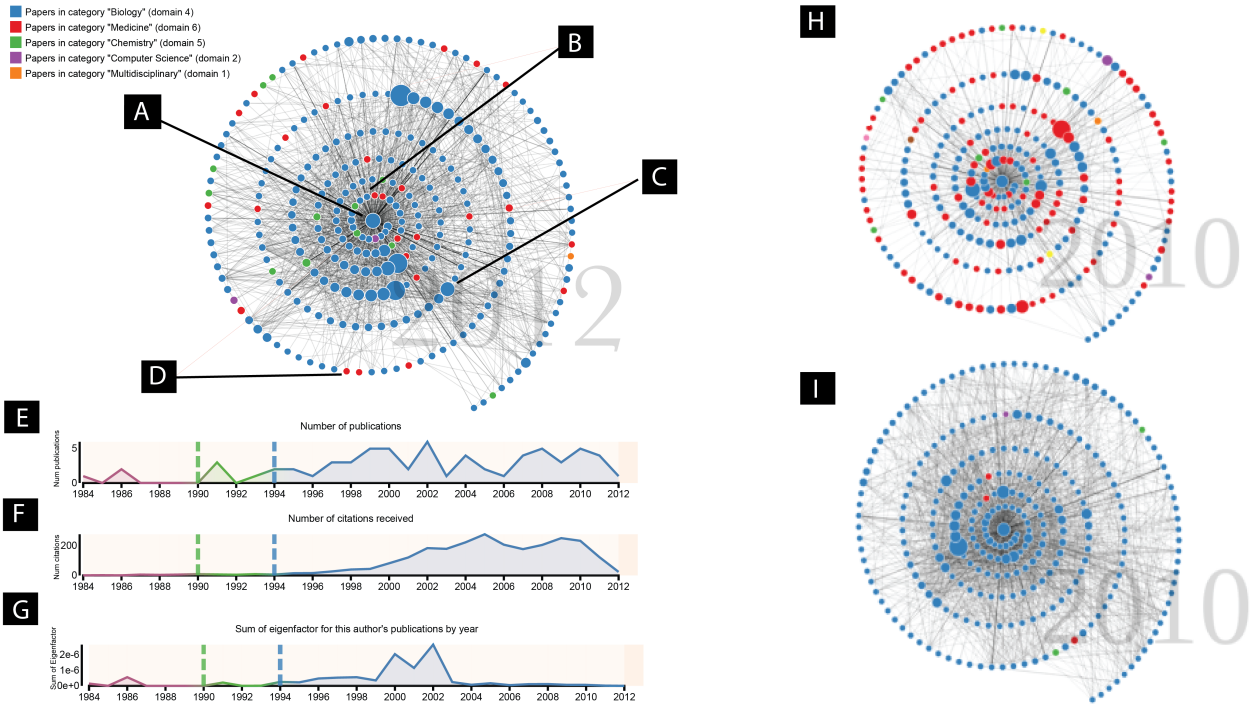

Figure 4.2: *Top Left:* (A) The center node represents all publications of a particular scholar. (B) Nodes that appear around the center represent publications that cited work by this scholar. (C) The size of the nodes show a citation-based indicator (Eigenfactor) of how influential that paper has been. (D) Colors show different fields to which the papers apply. *Bottom Left:* Integrated timeline charts below the network visualization. (E) Number of publications by the central scholar by year. (F) Number of citations received by the central scholar by year. (G) Sum of the Eigenfactor for all of the publications published by the central author in each year. Colors show the periods before, during, and after funding from the Pew program. *Right side:* Comparing the densities of two different graphs. (H) is a sparse graph that shows a diffuse influence across fields (i.e., interdisciplinary influence). (I) is a dense graph that shows a close-knit citation community within one domain.

Field of Study hierarchy. We choose a single domain based on the majority 2nd-level Field of Study assigned to each paper, concatenating multiple domains in the event of a tie.[3]

Graph representation: We represent scholarly publications as nodes in a network, and citations as directed links between them. Additional features relevant to assessing influence are stored as node attributes. These include year of publication, title, authors, and domain. The Eigenfactor score—a metric of influence for each paper that takes into account the number and quality of its citations, calculated across the entire data set—is also stored as an attribute of each node (see Background above).

We transform the data into a directed *egocentric network,* a subset of the total graph that considers one central node (the ego) and all of its neighbors (the alters), as well as all of the edges from alters to ego and between alters [36]. The center node represents the set of all papers authored by a particular scholar. This approach requires author identification as a subtask: determining which papers in a large scholarly database are authored by a given individual (see Implementation/Algorithm section below).

Taking this approach, the center (ego) node represents the total body of work authored by the scholar being visualized (Figure 4.2 A). All of the scholar's papers and their associated features are stored as attributes on the center node. All of the alter nodes represent individual papers that have cited any of the papers contained in the ego node. The alters all have at least one link to the center node—multiple if the paper cited more than one paper authored by the scholar of interest—as well as links to other papers that appear in the egocentric network. The Pew scholars we visualize have some variation in the total number of nodes in their network, typically between around 200 and 5000. As described below in Implementation/Algorithm, we limit the number of nodes displayed in the graph portion of the visualization (n=275 for all figures in this paper) in order to keep the level of visual complexity under control.

---

[3]The Field of Study designations for papers have been noted as a weakness of the current MAG data, with fields that are "dynamic, too specific, and field hierarchies are incoherent" [87]. As improvements are made to the underlying metadata, more effective visualizations can be generated.

Key Indicators over Time: Additional data transformations calculate key indicators of the scholar's career over time. Each of these indicators are calculated for each year: total number of publications authored by the scholar, total number of citations received by any of the scholar's papers, and sum of the Eigenfactor influence scores for all of the papers authored by the scholar in each year. Since we use the Eigenfactor score as a measure of the citation-based influence of an individual paper, the Eigenfactor score sum can be thought of as a measure of the total (citation-based) influence the scholar's output has had that year. Each of these indicators contextualizes the career-level data from a different angle. These indicators are visualized over time in linked timeline charts that appear below the graph display (See Figure 4.2–E, F, and G).

Validation: The most important data abstraction decision to validate is our conceptualization of influence. Through discussions with the Pew officers, and informed by a broader awareness of how influence can be conceptualized, we identified features in our data that reflected different facets of influence. Measures of citation counts and importance of publications in the overall network (i.e. Eigenfactor) show clear but rough indications of the influence a scholar has had. Features such as the domain of the citing work and the number of connections citing papers have had to other citing papers say something about the type of influence the scholar has had—whether it tends to be concentrated in a small community or diffuse to different areas.

Additional downstream validation of the data abstractions came through testing the design with Pew scholars (see Results).

Visual encoding and interaction design

The graph is represented visually in the common paradigm of the node-link diagram, with circular nodes representing vertices connected by straight lines representing the edges between them (Figure 4.2–A-D). The ego node, representing all of the central scholar's papers, is placed at the center of the display (A). The alter nodes, representing papers citing the ego's work, surround the ego node (B). All nodes and links are hidden initially, then are animated

in chronological order by year, extending in a spiral layout from the center node, beginning with the year of the earliest publication by the central scholar. A year counter behind the graph displays the publication year of the nodes and links currently appearing. The direction of the links is encoded by the animation: links are sent out from each citing node to the cited nodes after the citing node appears. The rate at which nodes appear is based on the number of nodes in each year, so that nodes appear more quickly in years with more citing; this is meant to lend more excitement to the more active years. If an alter node has more than one link to the center node (i.e. the paper has citations to multiple papers authored by the ego scholar), multiple lines are drawn on top of each other, so that edge thickness is mapped to number of citations. While the final node-link structure is often complex and interpreting the meaning of individual links is difficult, the intent is to convey a high-level view of the connections that form around a scholar in her citation community, and to allow relative comparisons of density. More focused analysis is supported by details on demand for a citing paper via mouseover of nodes. The user can also click on a node to be taken to either the full text of the paper (if available) or the paper's Microsoft Academic page.

Animation: The use of animation to show the network build over time was an important design choice throughout the process. A goal of this visualization was to use the data to tell a story that would be compelling to a wide audience. While it can have drawbacks, animation as a medium naturally draws attention and can encourage perceptions of narrative [134]. By using animation to encode time series data as observable changes, metaphoric change may also be communicated [169].

Spatial encoding: We experimented with multiple spatial encodings of the nodes. Initial designs used a force directed layout commonly used in node-link diagrams, to place the alter nodes around a fixed ego. This placement, however, tended to produce an overwhelming visual representation that was difficult to interpret (the "hairball" effect also commonly associated with node-link diagrams). It also did not make effective use of spatial placement as means of encoding something useful about the data.

We chose to place the nodes in a spiral pattern for several reasons. By ordering the nodes by year and placing them outward from the center, it allowed us to encode temporal information in the network—increased radial distance represents a more recent publication date, one that is later in the scholar's career. The original force-directed layout encoded publication date only temporally, with earlier dates being revealed earlier in the animation. The spiral layout adds the spatial encoding to reinforce this dimension, making the narrative easier to follow. Another advantage of the spiral placement was the ability to include more nodes in a limited space without too much overlap and confusion. The tradeoff of this placement is that it precludes optimizations intended to minimize edge crossing.

Other encodings: We chose to encode two additional features on the network's alter nodes: Eigenfactor and domain. The Eigenfactor of each paper is represented by the relative size of the node (Figure4.2 C). This allows the viewer to easily identify some of the most important papers that have cited the center scholar's work (see Background section above for more about Eigenfactor). The domain of each paper is represented by the color of the node (Figure4.2 D), with a legend generated for each scholar on the top left of the display identifying which colors map to which domains. The most common domain among the papers in the ego node is set as blue, and other domains that appear in the network are assigned to a categorical color scheme in order of frequency with which they appear in the graph. We chose a relative rather than absolute color scheme because there exist too many fields to assign each a color. In addition to showing individual papers from different fields, the extent of color variation in the total graph allows the viewer to see at a glance the extent to which the influence of a scholar's work tends to cross intellectual boundaries.

Timeline visualizations: Three timeline charts appear below the graph. The x-axis shows the years from the earliest paper authored by the center author to the last year in our data set for which we have data. Figure 4.2–E, F, and G show the timeline charts; see the Data abstraction section above for a description of the data abstractions shown. As the time progresses, the current year being visualized is highlighted in the timeline charts in orange. The years that have already been visualized are highlighted in faint orange. The

viewer may click on a year in the x-axis to move the animation forward or backward to the state of that year. One additional dimension was encoded for the interest of the Pew program officers—colors and vertical lines show the periods before, during, and after the funding that the Pew program provides to the scholars. This is one example of how overlaying additional data can help to add context to the overall story, and is discussed more in the Future Work section below.

Comparing visualizations: The scales used throughout the visualization—the mapping of Eigenfactor to node size, the color of the domain, and the y-axes on the timeline charts—are calculated relative to each individual scholar. This makes comparisons between different scholars on these dimensions difficult. This was a deliberate design choice. As discussed in Background above, quantitative metrics exist and are already widely used to compare scholars based on measures of output and citation counts. Our initial intent in working with the Pew program was to discourage comparison and ranking in favor of a more qualitative view of an individual scholar's influence. However, as we generated visualizations for different scholars, we did notice certain patterns that said something about the different types of influences. The right side of Figure 4.2 shows two different graphs, one dense and monochrome (H), the other sparse and colorful (I). One is not necessarily more influential than the other; rather, they exemplify two different ends of a spectrum of influence, which can be seen in the citation pattern around the scholars' work. The dense, monochrome graph shows a scholar who tends to have influence in a specific area, a close-knit group of researchers that have many connections to each other. The sparser, more colorful graph shows a scholar who has had diffuse influence in different disciplines. The papers that cite this type of scholar tend to cite other papers that appear in the graph less often, resulting in fewer links between alters and a sparser network. Supporting these types of comparisons will be important as we continue to develop these methods (see Future Work)

Implementation/Algorithm

Implementing the overall design is carried out in several stages: identifying the author in the database, collecting and caching the data, and drawing the visualization.

Author identification: Inaccurate *author disambiguation* is a threat to the validity of the depiction of scholarly impact. A unique identifier in the data set corresponds to an author identified by the collection algorithms; however, a single scholar may actually correspond to several IDs, and scholars with common names may be mistaken for different people due to inaccuracies in the algorithms Microsoft uses to collect the data. To mitigate the potentially misleading view of influence that can occur from disambiguation errors upon inputting only an author name, user input is required. The latest version of the system—hosted on `http://scholar.eigenfactor.org`, allows users to curate their own collections of papers, selecting and removing papers from the collections as they see fit before generating the data and visualization.[4]

Obtaining and Representing Data: The next stage of implementation is putting the data (stored in a MySQL database) into a network structure using the Python packages pandas [113] and NetworkX [79]. Starting with a graph with the ego node representing a scholar, the total set of papers associated with this scholar (as curated by a user) are stored as an attribute of the ego node. For each of these papers, the citing papers are collected and added to the graph as an alter node. Finally, for each citation by an alter paper, an edge is created between alter and ego if the cited paper is in the ego node, or between alter and alter if the cited paper appears in the graph.

Visualization Rendering: The final stage of implementation is the visualization, implemented using the open-source JavaScript library D3 [27]. For the network representation, in order to reduce the visual complexity, the number of total nodes is capped at 275; if there are more, the alter nodes are chosen based on Eigenfactor and whether they have associated domain information. The alter nodes are then sorted by year, placed in a spiral formation

[4]This site has been partially decommissioned, but a demo remains.

around the center, and hidden. The speed at which nodes appear is calculated based on the number of nodes in the current year being animated, using a threshold-based scale that sets the total time per year. This scale is set to achieve a balance between smooth narrative and having nodes appear faster in years with more activity. Years with very few nodes take .8 seconds, while years with 30 or more nodes take 4 seconds to animate (with multiple threshold settings in between); empty years take .3 seconds.

The number of nodes to visualize (n=275) and the spacing of the nodes is hard-coded, and was arrived at after some trial and error. The goal was to show as many nodes as possible in the space typically afforded by a web visualization, while avoiding excessive overplotting and occlusion. We arrived at this design after going through several iterations in collaboration with the Pew officers.

4.1.7 Results

Evaluation with Scholars

The Pew program held a three day meeting in November 2015 for their 30th reunion with approximately 400 scholars attending, ranging from the first class of 1985 to the class of 2011 (a scholar's Pew class is the year that he or she was accepted to the program and began to receive the four years of funding). Throughout the three days, the scholars attended research talks and social events. We set up a table with a display so that scholars could view and interact with the visualization during their down time. When a scholar approached the table, we demonstrated the visualization with her data and allowed her to watch and interact. We then asked open-ended questions to prompt a dialog.

During the reunion, we demonstrated the visualization with 26 scholars. We also allowed the scholars to access the visualization on their own online, and encouraged them to contact us with any feedback. Since the demonstrations, we have received approximately 20 emails and engaged in 15 informal conversations providing additional feedback. In this section, we discuss high-level observations that emerged from these demonstrations and conversations.

In the next section, we present several interesting individual stories that came out of the experience.

While interest in viewing and interacting with the visualization was high, many of the scholars approached with skepticism. Many scholars are wary of the limitations of evaluations based solely on publications, and a common frustration expressed among the Pew scholar was the use of measures such as citation counts and h-index. However, we observed that for most scholars reactions shifted to a generally positive tone after trying out the visualization. While the concerns were not completely assuaged, we believe that the scholars tended to appreciate how the visualization represented different dimensions of influence to present a richer picture than these common metrics. Several scholars noted this aloud.

Several scholars struggled with the fact that nodes represented citing papers, rather than the scholar's own papers. We suspect that this difficulty adjusting to nodes representing citing papers may be partly due to the emphasis on the individual scholar's papers in many current scholarly databases that offer ego-views, such as individual scholar's DBLP or Google Scholar profiles. An interesting avenue for future work is to integrate a depiction of the scholar's own papers as part of the visualization (see [145] for a step in this direction).

Another common issue with the data abstractions that came up during these validations was that of the difference between review articles and original research. Review articles tend to be highly cited papers, especially in the biomedical field, and thus may be overrepresented in the graph display. When the scholars interacted with the visualization and began identifying some of the larger nodes, they found that many of them were in fact review articles. While many scholars agreed that it was noteworthy to be cited by a prominent review article, some thought that review articles represented something different from original research, and thought that the distinction should be made clear. These comments made us aware that the influence of review articles can be a contentious topic among some scholars, who believe that they should be omitted entirely from influence measures. Future work can focus on making this distinction clearer, and devising ways of identifying which papers in the network scientists tend to consider more important and influential.

Most of the scholars were interested in comparing their data to others', asking tentative questions along the lines of, "Is my spiral good?" As discussed above in the Visual Encoding section, we tried to discourage these sorts of comparisons. While it is possible to see absolute differences between scholars–for example, by examining the scales in the y-axes of the timeline charts to see who had more publications or citations or comparing the density of the link structure across graphs (Figure 4.2 H, I)—the visualization is not designed to make these differences prominent. Our intent was to use our data to highlight the different types of influence these researchers have had, and it was usually possible, with some effort, to steer the focus in that direction.

Stories from the Scholars

One of the most interesting results to come out of the demonstrations was that viewing their data frequently prompted the scholars to reflect on their careers and to tell stories about how what they saw on the screen matched up with how they saw their own histories. There were many comments about how certain peaks or dips in the timeline charts, or changes in activity or color on the graph, corresponded to career shifts, restructuring of laboratories, or even significant personal events. The visualization thus served as a catalyst for communication around a particular scholar's trajectory, at some points fostering discovery by the scholar of influences and dynamics in their career of which they had not been aware. In this section, we present several specific stories that emerged.

One scholar, when shown her citation network, noticed that she had been cited by a prominent paper in the Agricultural Sciences literature (Figure 4.3A). At first she identified this as an error in the data. Her area of study is the cellular mechanisms underlying heart attack, and she didn't see herself as having any connection to the study of agriculture. However, on further reflection, she made the connection that a particular protein to which she had devoted a period of her career was also involved in meat tenderization. In this case, the self-reflection enabled by visualizing this scholar's citation network enabled her to identify an influence she had had on a completely different field, one which she hadn't considered before.

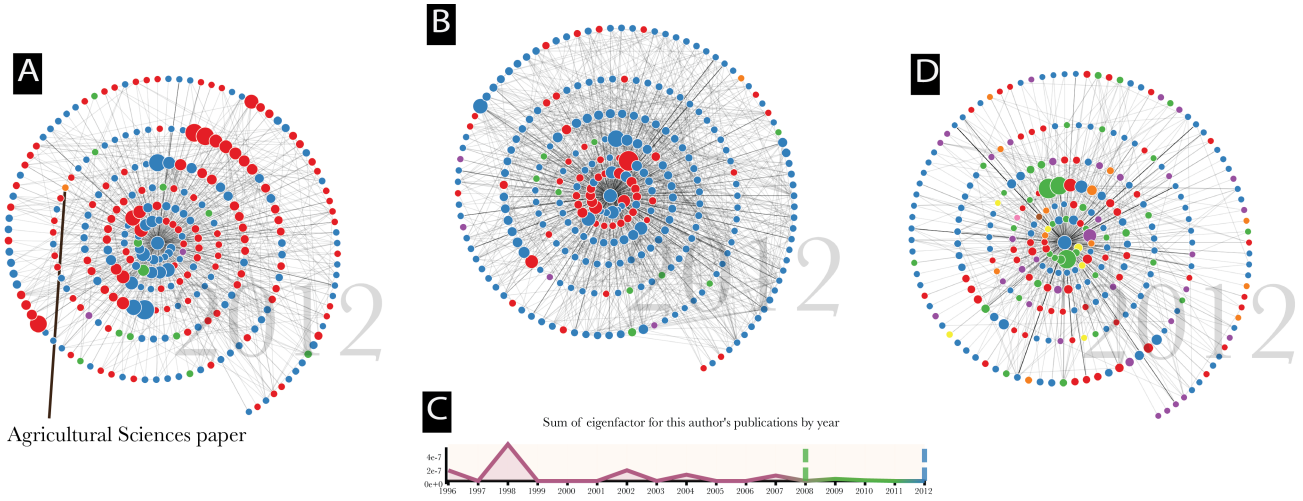

Figure 4.3: Four stories that emerged from demonstrations with the scholars. A) shows a scholar who had influence in a field she hadn't expected. B) shows a career shift reflected in changing color bands in the graph. C) shows an early-career peak in influence that prompted a scholar to reflect on the freedoms afforded by different research positions. D) shows a scholar with influences in very diverse areas.

Another scholar's graph showed a dramatic shift in color from the inner to the outer nodes of the spiral (Figure 4.3B). Talking to this scholar, he agreed that this reflected a major shift in his career, when the focus of his research changed from a topic in chemistry to one in biology. The papers that tended to cite his work changed as well, and the color patterns in his graph conveyed this shift in influence in a way that was easily recognizable to him. We also saw these shifts in color when scholars changed model organisms (e.g., *Arabidopsis* to *Drosophila*).

These methods do not tend to work as well for young scholars, as a longer career provides more input data for telling the story of their developing influence. Nevertheless, one young researcher noticed a peak in her Eigenfactor timeline chart that corresponded to some of her work in graduate school (Figure 4.3C). This led her to reflect on her time in graduate school and the boundary-pushing research that she conducated at that time. Her current research position, she said, allows much less of this type of freedom. In this case, the scholar was able

to imbue the data visualized on her chart with her personal story of how she felt about her research's ability to have meaningful influence.

One scholar, before viewing his visualization, jokingly commented on how unfocused he was—he tended to publish on a wide range of topics and expected his citation graph to reflect this. As promised, his graph turned out to be the most colorful we had seen, reflecting a career whose influence had reached researchers in chemistry, medicine, biology, material science, engineering, physics, computer science, and environmental science (Figure 4.3D). The alter nodes in his graph do not have many connections to each other, which is another indicator that his influence has reached a diverse set of communities. What this scholar referred to as "unfocused" scholarship could also be seen as diverse and boundary-spanning. Recent work has suggested that top-performing researchers form co-authorship networks that tend to be well connected and structured for bridging structural holes [178]; in this light, this interdisciplinary researcher may share characteristics with some of the world's most impactful scientists. This scholar enjoyed seeing his story visualized in this way, and wanted to feature the graph on his personal website. He also used the visualization as a chance to reflect on his future plans, mentioning that he expects the graph to get "even worse"—i.e. more colorful and reflective of more diffuse influence—in the future as citations stemming from his recent work increase.

4.1.8 Discussion and future work

While our focus in this work was developing narrative visualizations for the Pew scholars, we have already begun to use it to generate visualizations of scholars outside of the program. We have also applied the methods to entire fields of study rather than individual scholars [138]. As we continue in this direction, future work will address the generalizability of all of our design choices—whether, for example, it might be better to use the data to choose the proper number or spacing of nodes, rather than having this predetermined. Identifying which nodes to display is also a question we will revisit, as we reconsider which papers in a scholar's

network are most salient to show influence. To do this, we will ask the Pew scholars themselves to note their most influential papers. This was an idea from one of the Pew scholars.

Our goals in working with the Pew program centered around creating visualizations to help the program reflect on its set of highly influential scholars. This shifted our focus away from direct comparisons of different scholars. As we broaden our scope and generalize out to include scholars outside of the program, however, one of the most important directions for future work will involve turning more toward comparison—addressing the question of how to place one author's story in a larger context. Displaying two visualizations side by side would be one option, with an author's display appearing next to an appropriate control. Thought needs to go into selecting these controls—for example, an aggregated representation of other authors with similar careers, or, in the context of evaluating impact for funding agencies, an aggregated control based on scholars who applied for funding but were not awarded or did not accept funds.

Another direction for future work relates to the narrative nature of the visualization: how to incorporate different types of data into the story told by the animation. We have shown one example of overlaying additional data to deepen the context: the coloring and labeling of the timeline charts by Pew funding period. Adding this dimension helped the Pew officers and scholars to reflect on the stories and consider the effect that entering into and receiving funds from this program may have had. Other additional encodings could support program and individual evaluation in a number of other settings.

Other forms of data could also be integrated to further emphasize the visualization as a storytelling device. Automated annotation of salient shifts in the magnitude or domain of influence could help guide a novice user's interpretation. Multimedia storytelling through the integration of audio is another interesting avenue for future work. The Pew program, for example, has conducted interviews with most of its scholars and has both audio and transcripts available. Excerpts from these interviews, played at the proper time during the animation, could provide additional dimensions to the overall story of the scholar's career.

Expanding out from the case study with the Pew scholars, the website `http://scholar.eigenfactor.org` will serve as a launching pad to offer as a free service this and other tools to analyze and visualize scholarly influence using citation graphs. User data and feedback will be helpful in expanding and developing these tools.

4.1.9 Conclusion

We presented a design study in the domain of visualizing scholarly influence to tell a scholar's story, collaborating with the Pew Biomedical Scholars program and using their scholars as an initial case study. We described our design process of choosing data abstractions and visual encoding techniques in collaboration with Pew program officers, and detailed the implementation. We demonstrated the visualization with the scholars, and identified general trends and specific stories that showed how the visualization helped the scholars to reflect on their own influence. Finally, to generalize the methods to more scholars, we implemented a system which allows users to curate collections of papers and generate visualizations themselves.

4.1.10 Acknowledgements

We would like to thank the Pew Charitable Trust and the Chemical Heritage Foundation for funding, and for allowing us to interact with the program managers and the Pew Biomedical Scholars. We also want to thank Microsoft Research for providing the citation data through their Microsoft Academic Graph.

4.2 Case Studies in Science Visualization

In addition to the collaboration with the Pew scholarship program, a number of other organizations and individuals have expressed their interest in working with me to build interactive visualizations to explore and gain insight from the scholarly literature. This has provided me a host of opportunities to tackle this challenge from different angles. In this section I will describe a few of these projects, which together have resulted in my designing a suite of visualizations that represent the literature in different ways. The breadth of experience that this work has given me has been instrumental in informing my current ongoing work on SciSight, which I will discuss in the next section.

The common thread through all of these projects is that they are visual representations of collections of papers, and the relationships between them. In general, the problem of information overload can be mitigated by constraining the set of publications (as Autoreview does). However, in trying to understand the relationships between these papers—citation relations, for example—even a small set of papers can become complex and overwhelming. Translating these relationships into visual representations can make use of the human brain's ability to quickly identify visual patterns, which can lead to new insights about data [104]. The visualizations I describe here were each designed to represent these collections and their relationships in slightly different ways, to focus on different aspects of the data and different questions. Some of these projects did not lead to peer-reviewed publication, but all of these tools are published as open source projects.

The **cluster network diagram** (Figure 4.4 top left) is meant to show the interaction between two different fields based on the citations connecting those fields. I designed this tool as part of a collaboration with the researchers Megan Finn and Quinn Dupont, who were studying the interaction between research communities in Information Security (InfoSec) and Ethics. After identifying two collections of papers to represent the two different fields, we wanted to explore the research communities involved in the two fields, and to what extent they did or did not interact. I used citation-based communities identified by the Infomap

community detection algorithm. I visualized these communities as nodes in a network diagram. To represent within-community mix between the two fields (InfoSec and Ethics), I colored the nodes based on the ratio of papers in each field. To represent between-cluster interaction, I drew links weighted based on the number of citations between the clusters. The visualization is interactive: tooltips on the nodes give more information, and the user is able to toggle between different modes for the size of the nodes, representing either total number of papers, or the relative number of papers in either of the two fields. Using this visualization, we were able to identify two InfoSec clusters with very different apparent relationships to the Ethics community: one cluster with papers about botnets did seem to interact with the ethicists, while another with papers about security on the Android mobile platform was largely disconnected from the Ethics clusters. I wrote a short demo paper describing the tool [139].

While the cluster comparison diagram shows relationships between clusters of papers, the **interactive coauthorship network**, shown in Figure 4.4 (top right), seeks to show research communities in a different way, by showing social relationships between authors. In this representation, nodes represent the authors and links represent coauthorship collaborations. I built this visualization as part of my general exam, to show the relationships between groups of researchers who were publishing about ways to visualize network clusterings.[5] I used it as a tool to help surface the different groups of researchers working together on these topics. In this instance, communities emerge naturally from the structure of the network, as the physical simulation that pulls and pushes the author nodes according to their connections tends to separate the communities (labs in this case).

As part of a project with JSTOR, I developed two additional interactive visualizations to explore collections of papers. Figure 4.4 (bottom left) shows screenshots of these. One is a timeline view of all of the papers in the collection by year. The other is a network visualization that shows citation relationships between the papers in the collection, allowing

[5]Source code for this visualization is available at https://github.com/h1-the-swan/nodelink_vis_coauthorship

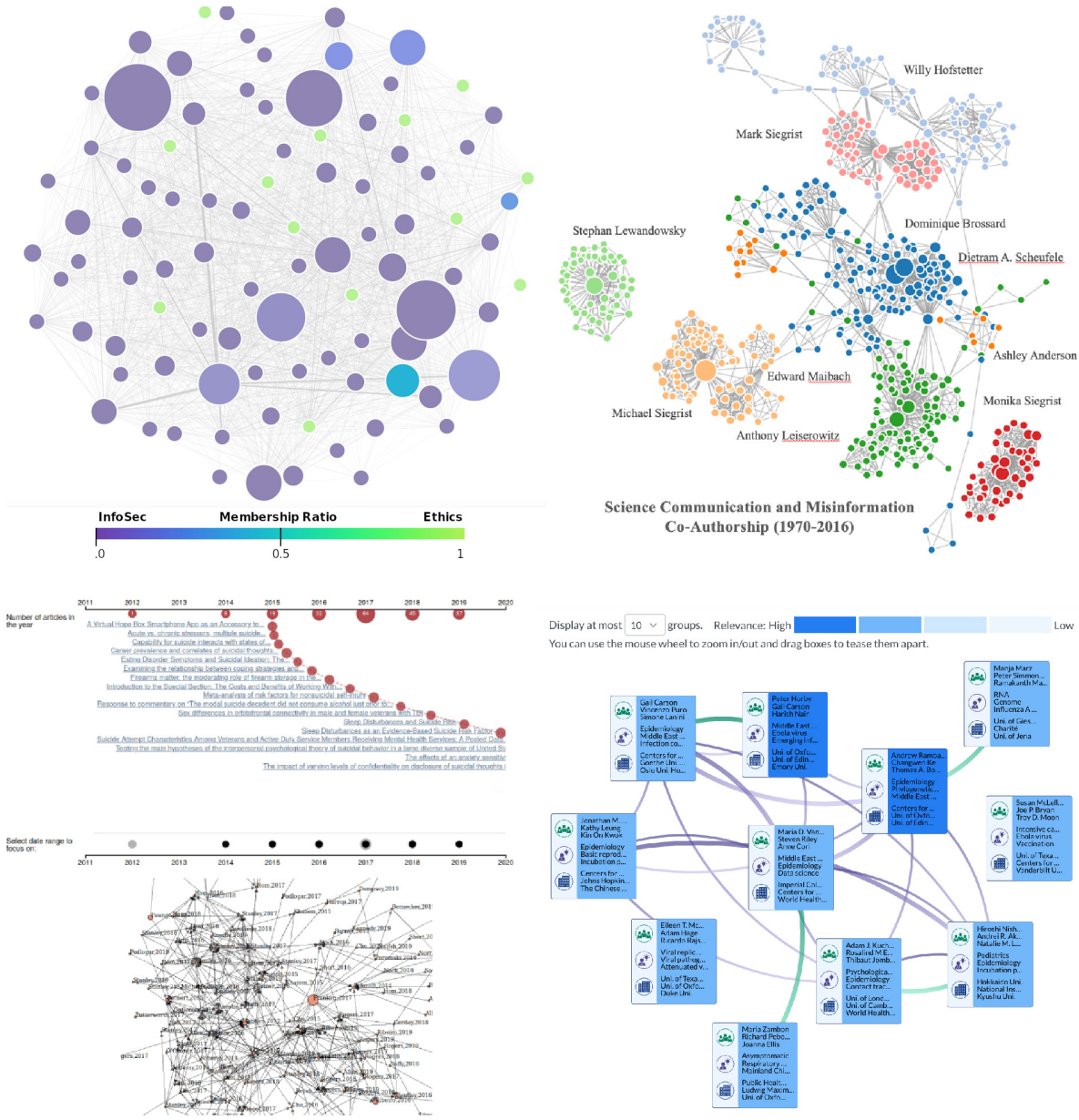

Figure 4.4: Visualizations for collections of papers. **Top left:** The **cluster network diagram** shows citation relationships between clusters of papers related to Information Security and Ethics. Clusters are colored according to the ratio of InfoSec or Ethics papers within, and links show citations between the clusters. **Top right: Coauthorship network** for researchers publishing in the fields of science communication and misinformation. Nodes represent authors; links represent joint authorship on the same paper. The colored clusters often correspond to research labs or groups. **Bottom left:** Interactive visualizations showing collections of articles: a timeline of papers by year (above) and a citation network (below). **Bottom right:** Screenshot of the **SciSight** visualization for COVID-19 research. The nodes of the network are "cards" representing groups of researchers, and links represent different types of relationships between them.

for focus on individual papers and chains of citation.[6] These were designed to show compact overviews of collections of papers representing JSTOR topics. The project allowed me to explore the temporal and relational aspects of paper collections. I am currently applying what I learned from this experience to current projects, including SciSight, discussed below.

The visualizations I have developed are meant to explore collections of papers from different angles, highlighting different patterns in the relationships between them. My work on these projects has given me experience and insight which I have brought to my current work on SciSight, in which I apply visualization techniques to show connections and gaps between researchers and groups of researchers. I describe this work in the following section.

4.3 SciSight / Bridger

4.3.1 Author Preface

The process of doing science tends to be organized into groups of researchers working on similar problems. This system, despite all its advantages, has a risk of suffering from information silos, when these research groups face barriers to communicating efficiently [175]. Cross-fertilization between these groups is important for driving innovation and advancing research [83, 98]. Gaps in the network of science—disconnected groups of researchers who might benefit from communicating—represent opportunities to assist in scientific advancement. This relates to Burt's theory of structural holes [34]: as he states, "Brokerage across the structural holes between groups provides a vision of options otherwise unseen" [33]. Because gaps are an absence of connection, they are often invisible to those involved. SciSight is a project aimed at identifying and discovering these gaps, by allowing a visual exploration of groups and their different dimensions of similarities and dissimilarities. The contributions of this project are in both tools and methods: the interactive visualization, as well as the methods of characterizing similarities between authors and groups along different dimensions.

---

[6] Code available at https://github.com/h1-the-swan/d3-article-timeline and https://github.com/h1-the-swan/d3-article-citations

SciSight (Figure 4.4 bottom left) is an ongoing collaboration with Tom Hope and others at the Allen Institute for Artificial Intelligence (AI2) [85]. This interactive visualization began as a tool to aid in learning about the groups of researchers working on the COVID-19 pandemic, and quickly began to receive some encouraging attention from high-profile sources [29, 88]. A paper on this tool has recently been accepted to the demo track at *EMNLP*, a leading conference in natural language processing. In the last few months, it has evolved into a larger project to explore groups of researchers and gaps between them.[7] Groups are characterized according to various dimensions, such as the tasks they work on, the methods and data sources they use, and the literature they tend to cite. Then, gaps between groups are identified as dissimilarities along some of these dimensions, where we might expect them to align. Figure 4.5 shows an example of this new direction, in which cards representing authors are presented to explore similarity in terms of methods and tasks.

In many ways, this project is a synthesis of the themes I have been working on throughout my graduate career. It has a strong focus on the social component of science. This relates to a project I have worked on previously analyzing cultural gaps between research communities by quantifying the jargon they use [175, 192]. This and other projects have informed my view of science as a social process, and SciSight continues this by exploring groups of collaborating researchers and quantifying different types of relationships between them, such as whether they share methods or tasks, whether they collaborate on papers, and whether they tend to cite the same prior work. SciSight also continues the theme of bringing together network science and natural language processing techniques to better understand science. This is a theme that features strongly in Autoreview, in which I use both of these types of features to identify related papers. It is also a theme through my internships at Meta and AI2. Both of these organizations have a strong focus on NLP, and I saw a great opportunity to bring in a networks perspective. It is also, as discussed in the previous section, a synthesis and a

[7] We are currently focusing on the domain of Computer Science, in part because our familiarity with it makes evaluation easier, but the methods and tools we are developing are general-purpose and will be applied to other domains, including COVID-19 research.

continuation of my previous projects developing visualizations to explore collections of papers to identify patterns in the scientific literature.

SciSight is an ambitious project with many different parts, and we have already made a lot of progress in a short amount of time. I have made a number of primary contributions to the project. I identified a corpus of 12 million computer science research papers, using data from MAG and AI2's Semantic Scholar (S2). Next, I identified overlapping communities of authors from the co-authorship network of these papers. In order to characterize papers by the different types of terms they mention, I used DyGIE++, a pre-trained deep learning model, to extract words and phrases from the papers' titles and abstracts identified as one of: method, task, material, or metric [177]. Using another deep learning model[8], I generated embeddings for each of these terms. I then adapted the SciSight visualization we had begun developing to show authors and groups together, finding similarities and differences between them using these embeddings (Figure 4.5).

Our next steps for this project will address something that has so far been missing in my visualization work—explicit evaluations with users to determine how well these tools are working. Evidence of the success of these tools have so far been in the form of more informal discussions with users, such as with the Pew Biomedical Scholars where we were able to find interesting stories by talking with the scholars as they played with data representing their careers and influence. The adoption of some of these tools by various organizations, such as Pew, HICSS, and the National Academy of Sciences, is further evidence of the tools' utility. Nevertheless, having more rigorous evaluations with users in which my methods are compared against baselines would provide a more compelling case. We recently completed these user studies, and have written our findings and submitted the paper to the CIKM conference. It is currently under review (as of June 21, 2021). This paper is included below as the rest of

[8]The model used to create embeddings for terms was CS-RoBERTa [77], which I fine-tuned for the task of semantic similarity using the Sentence-BERT framework [150].

section 4.3 [146].[9] (We recently changed the name of this part of the project from SciSight to Bridger.)

In addition to the user studies, we are continuing to build out the project in several ways. We will incorporate more dimensions of (dis)similarity, such as citation patterns. We will run analyses across the full corpus to study gaps at a large scale. And we will work toward scaling up the tool to be able to integrate it into a system like Semantic Scholar, and toward leveraging the methods in recommendation systems.

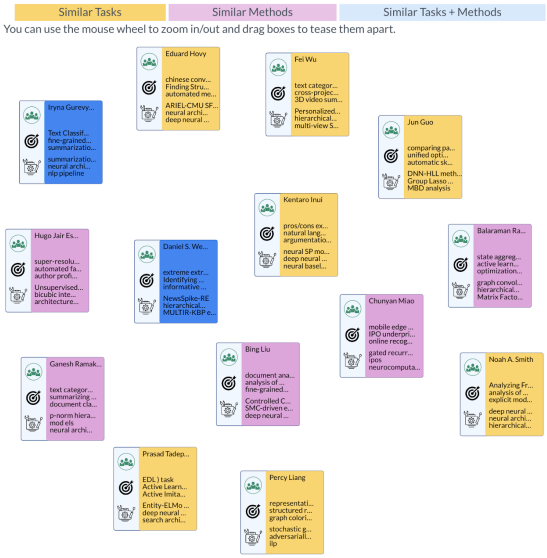

Figure 4.5: Screenshot of the SciSight visualization for computer science researchers. Cards represent individual authors, and colors show similarities among researchers based on the methods and tasks they use in their research.

[9]This study is currently under review. To cite material from section 4.3 of this chapter, please cite the original work: Jason Portenoy et al. "Bridger: Toward Bursting Scientific Filter Bubbles via Novel Author Discovery." In: *In review.* 2021

PUBLISHED WORK BEGINS HERE (IN REVIEW)

4.3.2 Abstract

Scientific silos can hinder innovation. These information "filter bubbles" and the growing challenge of information overload limit awareness across the literature, making it difficult to keep track of even narrow areas of interest, let alone discover new ones. Algorithmic curation and recommendation, which often prioritize relevance, can further reinforce these bubbles. In response, we describe Bridger, a system for facilitating discovery of scholars and their work, to explore design tradeoffs among relevant and novel recommendations. We construct a faceted representation of authors using information extracted from their papers and inferred personas. We explore approaches both for recommending new content and for *displaying* it in a manner that helps researchers to understand the work of authors who they are unfamiliar with. In studies with computer science researchers, our approach substantially improves users' abilities to do so. We develop an approach that locates *commonalities and contrasts* between scientists—retrieving partially similar authors, rather than aiming for strict similarity. We find this approach helps users discover authors useful for generating novel research ideas of relevance to their work, at a higher rate than a state-of-art neural model. Our analysis reveals that Bridger connects authors who have different citation profiles, publish in different venues, and are more distant in social co-authorship networks, raising the prospect of bridging diverse communities and facilitating discovery.

4.3.3 Introduction

> "Opinion and behavior are more homogeneous within than between groups... Brokerage across structural holes provides a vision of options otherwise unseen." (Burt, 2004)

The volume of papers in computer science continues to sky-rocket, with the DBLP computer science bibliography listing over 500,000 papers coming out in the year 2020

alone. [10] In particular, the field of AI has seen a meteoric growth in recent years, with new authors entering the field every hour [165]. Research scientists rely largely on search and recommendation services like Google Scholar and Semantic Scholar to keep pace with the growing literature and the authors who contribute to it. The literature retrieval services algorithmically decide what information to serve to scientists [17, 48], using information such as citations, textual content and clickthrough data, which inform machine learning models that output lists of ranked papers or authors.

In addition to the content of papers, these services rely on user behavior and queries. They adapt and reflect human input and, in turn, influence subsequent search behavior. This cycle of input, updating, engagement, and response can lead to an amplification of biases around searchers' prior awareness and knowledge [96]. Such biases include selective exposure [70], homophily [114], and the aversion to information from novel domains that require more cognitive effort to consider [83, 98]. By reinforcing these tendencies, algorithmic systems that filter and rank information run the risk of engendering so-called *filter bubbles* [135] that fail to show users novel content outside their narrower field of interest. How to mitigate these effects is a challenging open question for algorithmic recommendation systems [127], explored recently for movies [210] and in e-commerce [73] by surfacing serendipitous content that is aimed at being both novel and relevant (see §4.3.4).

Scientific filter bubbles can be costly to individual researchers and for the evolution of science as a whole. They may lead scientists to concentrate on narrower niches [100], reinforcing citation inequality and bias [128] and limiting cross-fertilization among different areas that could catalyze innovation [83]. Addressing filter bubbles in general, in domains such as social media and e-commerce recommendations, is a hard and unsolved problem [73, 45, 210]. The problem is especially difficult in the scientific domain. The scientific literature consists of complex models and theories, specialized language, and an endless diversity of

[10] `https://dblp.org/statistics/publicationsperyear.html`

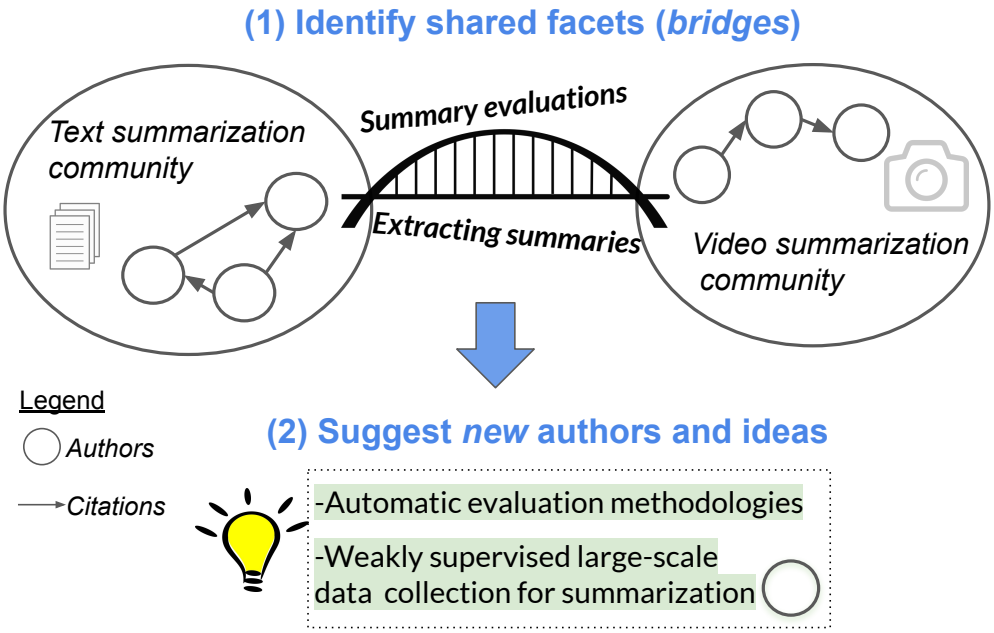

Figure 4.6: *Bursting scientific bubbles with Bridger.* The overarching goal is to (1) find commonalities among authors working in different areas and unaware of one another, and (2) suggest novel and valuable authors and their work, unlikely discovered otherwise due to their disparities.

continuously emerging concepts. Connecting blindly across these cultural boundaries requires significant cognitive effort [175].

Our goal is to connect scientists across the literature to less familiar, but related authors and ideas, and thereby facilitate scientific discovery. Working toward this goal, we developed Bridger, which is illustrated in Figure 4.6. The Bridger system employs a faceted, multidimensional author representation that includes information that is extracted automatically from papers, including tasks, methods and resources, and automatically inferred *personas*. The personas reflect the different focus areas on which each scientist works. Each of these aspects is embedded in a vector space based on its content, allowing the system to identify authors with commonalities along specific dimensions and not others, such as authors working on similar tasks but not using similar methods.

We explore the utility of this representation in experiments with computer science researchers. For the task of discovering new and useful authors, we find that Bridger's approach does better than a strong neural model currently employed by a public scholarly search engine for serving recommendations—despite Bridger's focus on surfacing novel areas and authors, and the built-in biases associated with this novelty. In addition to assessing *what* content, we also consider *how* to display it in a way that enables users to rapidly understand what new authors work on. We employ Bridger as an experimental platform to explore which facets are displayed to users, investigating various design choices and tradeoffs between relevance and novelty. We obtain substantially better results in terms of user understanding of profiles of unknown authors, when displaying information taken from our author representation.

Finally, we conduct in-depth analyses that show that Bridger surfaces novel and valuable authors and their work, unlikely to be discovered in the absence of Bridger due to publication in different publication venues, citing and being cited by non-overlapping communities, and having greater distances in the social co-authorship network. Taken together, the ability to uncover novel and useful authors and ideas, and to serve this information to users in an effective and intuitive manner, suggests a future where automated systems are put to work to build bridges across communities, rather than blindly reinforcing existing filter bubbles.

4.3.4 Related Work

Filter Bubbles and Recommendations Filter bubbles and related biases have been studied in the context of recommender systems [127], with recent work studying e-commerce websites [73] and widely-used algorithms [210]. One approach that has been explored for mitigating these biases is judging recommendations not only by accuracy, but with other metrics such as diversity (difference between recommendations) [196, 45], novelty (items assumed unknown to the user) [208], and serendipity (a measure of relevance and surprise associated with a positive emotional response) [186]. The notion of serendipity is notoriously hard to quantitatively define and measure [92, 43, 186, 190]; recently, user studies have

explored human perceptions of serendipity [43, 186], yet this problem remains very much open. A distinct, novel feature of our work is the focus on the scientific domain, and that unlike the standard recommendation system setting we measure our system's utility in terms of boosting users' ability to discover authors that spur *new ideas* for research. In experiments with computer science researchers, we explore interventions that could potentially help provide bridges to authors working in diverse areas, with an approach based on finding faceted commonalities and contrasts between researchers.

Inspirational Stimuli Our work is related to a relatively small body of literature focused on computational tools for boosting creativity [83, 98, 74, 84]. Experiments in this area typically involve giving participants a specific concrete problem, and examining methods for helping them come up with creative solutions [83, 84]. In our efforts reported in this paper, we do not assume to be given a single concrete problem. Rather, we are given *authors* and their papers, and automatically identify useful and novel inspirations in the form of other authors and their contributions. These computationally complex objects are very different to the short, single snippets typically used in this line of work [83, 84]. A recurring theme in this area is the notion of a "sweet spot" for inspiration: not too similar to a given problem that a user aims to solve, and not too far afield. Finding such a sweet spot remains an important challenge. We study a related notion of balancing between commonalities and contrasts between researchers.

Scientific Recommendations Work in this area typically focuses on recommending *papers*, using proxies such as citations or co-authorship links in place of ground truth [164, 15, 141], or a combination of text and citation data [48]. In addition to being noisy proxies in terms of relevance, these signals reinforce existing patterns of citation or collaboration, and are not indicative of papers or authors that would help users generate *novel* research directions — the focus of Bridger. Furthermore, we perform controlled experiments with researchers to be able to better evaluate our approach without the biases involved in learning from

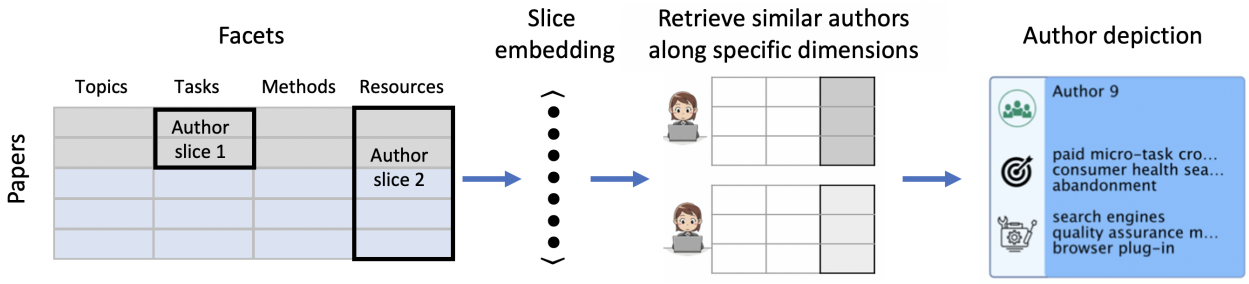

Figure 4.7: Overview of Bridger's author representation, retrieval, and depiction. Users are represented in terms of a matrix with rows corresponding to papers, and columns corresponding to facets. Bridger finds suggested authors who match along certain "slices" of the user's data – certain facets, subsets of papers, or both.

observational data on citations or co-authorship. One related recent direction considers the problem of diversifying social connections made between academic conference attendees [167, 168, 187], by definition a relatively narrow group working in closely-related areas, using attendee metadata or publication similarity.

4.3.5 Bridger: System Overview

In this section we present Bridger's methods and system. The Bridger system is designed to enable the study of different design choices for scientific author and idea discovery. We start by describing our representation for papers. We then describe how Bridger represents authors by aggregating paper-level information and clustering into *personas*. We use these representations to provide a general framework for discovery, based on author matching along specific dimensions (Figure 4.7). Finally, we describe depiction of author information, and end with implementation details.

Paper representations

Paper Information Each paper P contains rich, potentially useful information. This includes raw text such as in a paper's abstract, incoming and out-going citations, publication date, venues, and more. One key representation we derive from each paper P is a vector

representation $\tilde{P}$, using a state-of-art scientific paper embedding model. This neural model captures overall coarse-grained topical information on papers, shown to be powerful in clustering and retrieving papers [48].

Another key representation is based on fine-grained facets obtained from papers. Let $\mathcal{T}_{P_i} = \{t_1, t_2, \ldots\}$ be a set of *terms* appearing in paper i. Each term is associated with a specific *facet* (category). We consider several categories of terms in this paper: coarse-grained paper topics inferred from the text [183], and fine-grained spans of text referring to *methods, tasks and resources* — core aspects of computer science papers [38] — automatically extracted from paper i with a scientific named entity recognition model. [177] Each term t is located in a "cell" in the matrix illustrated in Figure 4.7, with facets corresponding to the columns and papers to rows. Each term $t \in \mathcal{T}_{P_i}$ is also embedded in a vector space using a neural language model (see §4.3.5), yielding a $\tilde{t}$ vector for each term.

Author representations

We represent an author, $\mathcal{A}$, as a set of *personas* in which each persona is encoded with facet-wide aggregations of term embeddings across a set of papers. Figure 4.7 illustrates this with outlines of "slices" in bold – subsets of rows and columns in the illustrated matrix, corresponding to personas (rows) and facets (columns).

Author personas Each author $\mathcal{A}$ can work in multiple areas. In our setting, this can be important for understanding the different interests of authors, enabling more control on author suggestions. We experiment with a clustering-based approach for constructing *personas*, $\mathrm{P}_{\mathcal{A}}$, based on inferring for each set of author papers $\mathcal{P}_{\mathcal{A}}$ a segmentation into K subsets reflecting a common theme — illustrated as subsets of rows in the matrix in Figure 4.7. We also experiment with a clustering based on the network of co-authorship collaborations $\mathcal{A}$ takes part in. See §4.3.5 for details on clustering. As discussed later (§4.3.6), we find that the former approach in which authors are represented with clusters of papers elicits considerably better feedback from scholars participating in our experiments.

Co-authorship information Each paper P is in practice authored by multiple people, i.e., it can belong to multiple authors $\mathcal{A}$. Each author assumes a *position* k for a given paper, potentially reflecting the strength of affinity to the paper. As discussed below (§4.3.5), we make use of this affinity in determining what weight to assign terms $\mathcal{T}_{P_i}$ for a given paper and given author.

Author-level facets Finally, using the above information on authors and their papers, we construct multiple author-level *facets* that capture different aggregate aspects of $\mathcal{A}$. More formally, in this paper we focus our experiments on author facets $\mathcal{V}_\mathcal{A} = \{\mathbf{m}, \mathbf{t}, \mathbf{r}\}$, where $\mathbf{m}$ is an aggregate embedding of $\mathcal{A}$'s *method* facets, $\mathbf{t}$ is an embedding capturing $\mathcal{A}$'s *tasks*, $\mathbf{r}$ represents $\mathcal{A}$'s *resources*. In addition, we also construct these facets separately for each one of the author's personas $\mathrm{P}_\mathcal{A}$ — corresponding to the "slice embeddings" illustrated in Figure 4.7. In analyses of our experimental results (§4.3.7), we also study other types of information such as citations and venues; we omit them from the formal notations to simplify presentation.

Approaches for recommending authors

For a given author $\mathcal{A}$ using Bridger, we are interested in automatically suggesting new authors working on areas that are relevant to $\mathcal{A}$ but also likely to be interesting and spark new ideas. More formally, we are given a user $\mathcal{A}$, their set of personas $\mathrm{P}_\mathcal{A}$, and for each persona its faceted representation $\mathcal{V}_\mathcal{A} = \{\mathbf{m}, \mathbf{t}, \mathbf{r}\}$. We are also given a large pool of authors across computer science, $\{\mathcal{A}_1, \mathcal{A}_2, \ldots\}$, from which we aim to retrieve author suggestions to show $\mathcal{A}$.

Baseline model We employ SPECTER [48], a strong neural model we compare to, trained to capture overall topical similarity between papers based on text and citation signals (see [48] for details) and used for serving recommendations as part of a large public academic search system. For each of author $\mathcal{A}$'s papers P, we use this neural model to obtain an embedding $\tilde{P}$. We then derive an aggregate author-level representation $\tilde{\mathbf{p}}$ (e.g., by weighted averaging that takes author-term affinity into account, see §4.3.5). Similar authors are computed using

a simple distance measure over the dense embedding space. As discussed in the introduction and §4.3.4, this approach focuses on retrieving authors with the most overall similar papers to $\mathcal{A}$. Intuitively, the baseline can be thought of as "summing over" both the rows and columns of the author matrix in Figure 4.7. By aggregating across all of $\mathcal{A}$'s papers, information on finer-grained sub-interests may be lost. In addition, by being trained on citation signals, it may be further biased and prone to favor highly-cited papers or authors.

To address these issues, we explore a formulation of the author discovery problem in terms of matching authors along specific dimensions that allow more fine-grained control – such as by using only a subset of views in $\mathcal{V}_{\mathcal{A}}$, or only a subset of $\mathcal{A}$'s papers, or both — as in the row and column *slices* seen in Figure 4.7. This decomposition of authors also enables us to explore *contrasts* along specific author dimensions, e.g., finding authors who use similar tasks to $\mathcal{A}$ but use very different methods or resources.

- **Single-facet matches** For each author $\mathcal{A}_i$ in the pool of authors $\{\mathcal{A}_1, \mathcal{A}_2, \ldots\}$, we obtain their respective aggregate representations $\mathcal{V}_{\mathcal{A}_i} = \{\mathbf{m}, \mathbf{t}, \mathbf{r}\}$. We then retrieve authors with similar embeddings to $\mathcal{A}$ along one dimension (or matrix columns in Figure 4.7; e.g., $\mathbf{r}$ for resources), ignoring the others. Unlike the baseline model, which aggregates *all* information appearing in $\mathcal{A}$'s papers – tasks, methods, resources, general topics, and any other textual information – this approach is able to disentangle *specific* aspects of an author, potentially enabling discovery of more novel, remote connections that can expose users to more diverse ideas and cross-fertilization opportunities.

- **Contrasts** Finding matches along *one* dimension does not guarantee retrieving authors who are *distant* along the others. As an example, finding authors working on *tasks* related to *scientific knowledge discovery* and *information extraction from texts*, could be authors who use a diverse range of *resources*, such as *scientific papers*, *clinical notes*, etc. While the immense diversity in scientific literature makes it likely that focusing on similarity along one dimension only will still surface diverse results in terms of the other (see results in §4.3.7), we seek to further ensure this.

To do so, we apply a simple approach inspired by recent work on retrieving inspirations [84]: We first retrieve the top K authors $\{\mathcal{A}_1, \mathcal{A}_2, \ldots \mathcal{A}_K\}$ that are most similar to $\mathcal{A}$ along one dimension (e.g., $\mathbf{t}$), for some relatively *large* K (e.g., $K = 1000$). We then rank this narrower list inversely by another dimension (e.g., $\mathbf{r}$), and show user $\mathcal{A}$ authors from the top of this list. Intuitively, this approach helps balance relevance and novelty by finding authors who are *similar* enough along one dimension, and within that subset find authors who are relatively *distant* along another.

- **Persona-based matching** Finally, to account for the different focus areas authors may have, instead of aggregating over *all* of an author's papers, we perform the same single-view and contrast-based retrieval using the author's personas $\mathrm{P}_{\mathcal{A}}$ — or, in other words, row-and-column slices of the matrix in Figure 4.7.

Depicting Recommended Authors

Our representation allows us to explore multiple design choices not only for *which* authors we show users, but also *how* we show them. In our experiments (§4.3.6, §4.3.7), we evaluate authors' facets and personas in terms of their utility for helping researchers learn about new authors, and for controlling how authors are filtered.

Term ranking algorithms to explain what authors work on Researchers, flooded with constant streams of papers, typically have a very limited attention span to consider whether some new author or piece of information is relevant to them. It is thus important that the information we display for each author (such as their main methods, tasks, resources, and also papers) is *ranked*, such that the most important or relevant terms appear first. We explore different approaches to rank the displayed terms, balancing between *relevance* (or centrality) of each term for a given author, and *coverage* over the various topics the author works on. We compare between several approaches, including a customized relevance metric

we design, in a user study with researchers (§4.3.6). We discuss in greater detail the ranking approaches we try in §4.3.5.

Retrieval explanations In addition to term ranking approaches aimed at explaining to users of Bridger what a new suggested author works on, we also provide users with two rankings that are geared for explaining how the retrieved authors relate to them. First, we allow users to rank author terms $\mathcal{T}$ by how similar they are to their own list of terms (for each facet, separately). Second, users can also rank each author's *papers* by how similar they are to their own — showing the most similar papers first. These explanations can be regarded as a kind of "anchor" for increasing trust, which could be especially important when suggesting novel, unfamiliar content.

Implementation details

Data We use data from the Microsoft Academic Graph (MAG) [160]. We use a snapshot of this dataset from March 1, 2021. We also link the papers in the dataset to those in an a large public academic search engine.[11] We limit the papers and associated entities to those designated as Computer Science papers. We focus on authors' recent work, limiting the papers to those published between 2015 and 2021, resulting in 4,650,474 papers from 6,433,064 authors. Despite using disambiguated MAG author data, we observe the challenge of author ambiguity still persists [163]. In our experiments, we thus exclude participants with very few papers (see §4.3.7), since disambiguation errors in their papers stand out prominently.

Term Extraction We extract terms (spans of text) referring to tasks, methods, and resources mentioned in paper abstracts and titles, using the state-of-art DyGIE++ IE model [177] trained on SciERC [108]. We extracted 10,445,233 tasks, 20,705,854 methods, and 4,978,748 resources from 3,594,975 papers. We also use MAG topics, higher-level coarse-

[11]Redacted for anonymity.

grained topics available for each paper in MAG. We expand abbreviations in the extracted terms using the algorithm in [157] implemented in ScispaCy [126].

Scoring papers by relevance to an author The papers published by an author have varying levels of importance with regard to that author's overall body of publications. To capture this, we use a simple heuristic that takes into account two factors: the author's position in a paper as a measure of affinity (see §4.3.5), and the paper's overall impact in terms of citations. More formally, for each author $\mathcal{A}$, we assign a weight $w_{\mathcal{A},P}$ to each paper P in $P_{\mathcal{A}}$, $w_{\mathcal{A},P} = \texttt{pos}_{\mathcal{A},P} \times \texttt{Rank}_P$, where $\texttt{pos}_{\mathcal{A},P}$ is 1.0 if $\mathcal{A}$ is first or last author on P and 0.75 otherwise, and $\texttt{Rank}_P$ is MAG's assigned paper Rank (a citation-based measure of importance, see [183] for details), normalized by min-max scaling to a value between .5 and 1.

Author similarity We explore several approaches for author similarity and retrieval, all based on paper-level aggregation as discussed in §4.3.5. For the document-level SPECTER baseline model discussed in §4.3.5, we obtain 768-dimensional embeddings for all of the papers. To determine similarity between authors, we take the average embedding of each author's papers, weighted by the paper relevance score described above. We then compute the cosine similarity between this author and the average embedding of every other author. For our faceted approach, we compute similarities along each authors' facets, using embeddings we create for each term in each facet. The model used to create embeddings was CS-RoBERTa [77], which we fine-tuned for the task of semantic similarity using the Sentence-BERT framework [150]. For each author or persona, we calculate an aggregate representation along each facet by taking the average embedding of the terms in all of the papers, weighted by the relevance score of each associated paper.

Identification of personas We infer author personas using two different approaches. For the first approach we cluster the co-authorship network using the ego-splitting framework in [62]. In a second approach, we cluster each authors' papers by their SPECTER embeddings

using agglomerative clustering with Ward linkage [122] on the Euclidean distances between embedding vectors.[12] In our user studies, we show participants their personas and the details of each one (papers, facets, etc.).[13] To make this manageable, we sort the clusters (personas) based on each cluster's most highly ranked paper according to MAG's assigned rank, and show participants only their top two personas.

Term ranking for Author Depiction We evaluate several different strategies to rank terms (methods, tasks, resources) shown to users in Experiment I (§4.3.6):

- **TextRank:** For each term t in an author's set of papers, we create a graph $G_F = (V, E)$ with vertices V the terms and weighted edges E, where weight w_{ij} is the euclidean distance between the embedding vectors $\tilde{t_i}$ and $\tilde{t_j}$. We score each term t_i according to its PageRank value in G_F [115].

- **TF-IDF** For each t, we compute TF-IDF across all authors, considering each author as a "document" (bag of terms) in the IDF (inverse document frequency) term, counting each term once per paper. We calculate the TF-IDF score for each term for each author, and use this as the term's score.

- **Author relevance score** For each t, we calculate the sum of the term's relevance scores (§ 4.3.5) derived from their associated papers. If a term is used in multiple papers, the associated paper's score is used for each summand.

- **Random** Each term t is assigned a random rank.

[12]Implemented in the scikit-learn Python library [136]. Distance threshold of 85.

[13]Some authors do not have detected personas; we observe this to often be the case with early-career researchers.

4.3.6 Experiment I: Author Depiction

In systems that help people find authors, such as Microsoft Academic Graph, Google Scholar, and AMiner [182], authors are often described in terms of a few high-level topics. In advance of exploring how we might leverage facets to engage researchers with a diverse set of authors, we performed a user study to gain a better understanding of what information might prove useful when depicting authors. We started from a base of Microsoft Academic Graph (MAG) topics, and then added their extracted facets (tasks, methods, resources). We investigated the following research questions:

- **RQ1**: Do tasks, methods, and/or resources complement MAG topics in depicting an author's research?
- **RQ2**: Which term ranking best reflects an author's interests?
- **RQ3**: Do tasks, methods, and/or resources complement MAG topics in helping users gain a better picture of the research interests of *unknown* authors?
- **RQ4**: Do personas well-reflect authors' different focus areas?

Experiment Design

Thirteen computer-science researchers were recruited for the experiment through Slack channels and mailing lists. Study sessions were one-hour, semi-structured interviews recorded over Zoom. The participants engaged in think-aloud throughout the study. They evaluated a depiction of a known author (e.g., research mentor) for accuracy in depicting their research, as well as depictions of five *unknown* authors for usefulness in learning about new authors. Throughout all parts of the experiment, the interviewer asked follow-up questions regarding the participant's think-aloud and reactions. To address **RQ1** and **RQ2**, the participants first evaluated the accuracy of a known author's depiction.

Step I. To begin, we presented the participant with only the top ten MAG topics for the known author. We asked them to mark any topic that was unclear, too generic, or did not reflect the author's research well. Next, we provided five more potential lists of terms. One of these lists consisted of the next 10 top topics. The other four presented 10 tasks, each selected as the top-10 ranked terms using the strategies described in §4.3.5. We asked participants to rank the five lists (as a whole) in terms of how well they complemented the first list (with an option to select none).

Step II. The process then repeated for five more potential lists to complement the original topics and the highest-ranked second list selected in Step I — this time, with methods instead of tasks. If the participant ranked a methods list highest, we then presented the participant with a resources list that used the ranking strategy preferred by the participant, and asked whether or not this list complemented those shown so far.

Step III. To address **RQ3**, participants next evaluate the utility of author depictions for five unknown authors. To describe each unknown author, we provided the participant topics, tasks, methods, and resources lists with 10 terms each. Each were ranked using tf-idf as a default. The participant noted whether or not each additional non-topics list complemented the preceding lists in helping them understand what kind of research the unknown author does.

Step IV. Finally, for **RQ4**, we asked participants to evaluate the known author's distinct personas presented in terms of tasks, which were ranked using tf-idf. On a Likert-type scale of 1-5, participants rated their agreement with the statement, "The personas reflect the author's different research interests (since the year 2015) well."

Results

Results for RQ1 The majority of participants found that tasks, methods, and resources complemented topics to describe a known author's research. For both tasks and methods, 11 of 13 participants felt that seeing information about that facet, more so than additional top MAG topics or no additional information, complemented the original

top ten MAG topics. The prevailing grievance with the additional MAG topics was that they were too general. Furthermore, 7 of 9 participants who evaluated a resources list thought that it complemented the preceding lists.

Results for RQ2 Participants overall preferred the relevance score ranking strategy for tasks and methods. We compared the four ranking strategies and MAG topics baseline strategy for both tasks and methods. For each participant, we awarded points to each strategy based on its position in the participant's ranking of the five strategies. We awarded the least favorite strategy one point and the most favorite strategy five points. Since there were 13 participants, a strategy could accumulate anywhere between 13 and 65 points. Separately, we counted how many times each strategy was a participant's favorite strategy (Figure 4.8c, d). With regards to tasks, TextRank and tf-idf accrued the most points from participants, with the relevance score trailing close behind (Figure 4.8a). Meanwhile, the MAG topics baseline accrued the least points, even fewer than the random task ranking strategy. In addition, relevance score and TextRank were chosen most often as the favorite task ranking strategy (Figure 4.8c). With regards to methods, the relevance score ranking strategy performed best in terms of both total points (Figure 4.8b) and favorite strategy (Figure 4.8d).

Results for RQ3 Participants generally found tasks, methods, and resources helpful to better understand what kind of research an unknown author does. To calculate how many participants were in favor of including tasks, methods, and resources to help them better understand an author, we determined the average of each participant's binary response per facet. Adding up the 13 responses for each facet, we saw that the majority of participants thought each additional facet helped them understand the unknown author better. All 13 participants found the tasks helpful, eight found the methods helpful, and 12 found the resources helpful. As an example, P12 connected an unknown author's topics, tasks, and methods to better understand them: "*I wouldn't have known they were an information*

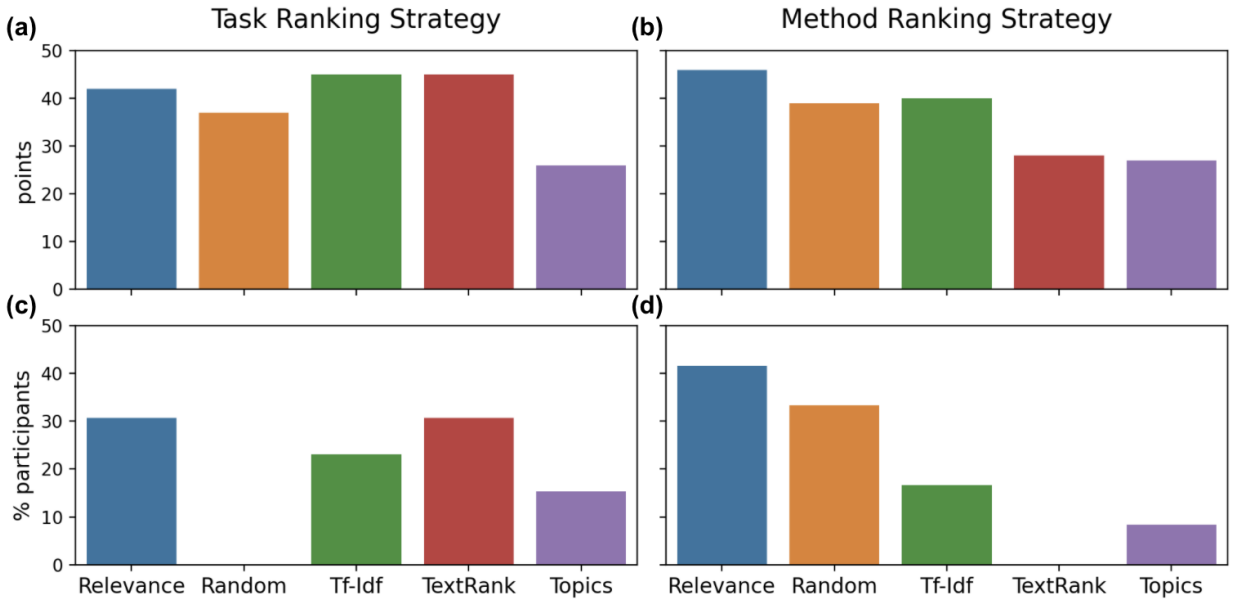

Figure 4.8: Points awarded to each ranking strategy for tasks (a) and methods (b), and percentage of participants who favored each strategy most for tasks (c) and methods (d).

retrieval person from the [topics] at all.... The previous things [in topics and tasks] that mentioned translation and information retrieval and kind of separately... This [methods section] connects the dots for me, which is nice." Interestingly, methods were not viewed to be as useful as tasks or resources. The majority of participants cited unfamiliar terms as a key issue.

Results for RQ4 Participants indicate preference for personas selected based on papers rather than co-authorship. After the experiment, six participants were informally asked to compare the experiment's personas selected based on co-authorship with the personas based on paper-based clustering (see §4.3.5). Four of them preferred the updated version. Furthermore, one of the users who preferred the old version still thought the updated version had better personas themselves and merely did not like the updated personas' ordering. In addition, all six participants liked seeing the personas in terms of papers. In our experiment

in §4.3.7, we observed much higher satisfaction with the updated personas in comparison to the original personas of this experiment.

4.3.7 Experiment II: Author Discovery

We now turn to our main experiment, exploring whether facets can be employed in Bridger to spur users to discover valuable and novel authors and their work. We use our two author-ranking strategies (§4.3.5), one based on similar tasks alone (`sT`) and the other on similar tasks with contrasting (distant) methods (`sTdM`). We compare these strategies to the SPECTER (`ss`) baseline. More specifically, we investigated the following research questions:

- **RQ5**: Do `sT` and `sTdM`, in comparison to SPECTER, surface suggestions of authors that are considered novel and valuable?
- **RQ6**: Does sorting based on personas help users find more novel and valuable author suggestions?

Experiment Design

Twenty computer-science researchers participated in the experiment after recruitment through Slack channels and mailing lists. All participants were shown results based on their overall papers (without personas) consisting of 12 author cards they evaluated one by one. Four cards were included for each of `sT`, `sTdM`, and `ss`. We only show cards for authors who are at least 2 hops away in the co-authorship graph from the user, filtering authors they had previously worked with.

For participants who had at least two associated personas, we also presented them with authors suggested based on each separate persona: four author cards for each of their top two personas (two under `sT` and two under `sTdM`). Whether the participants saw the personas before or after the non-persona part was randomized.

Each author card provides a detailed depiction of that author (see Figure 4.7). The author's name and affiliation is hidden in this experiment to mitigate bias. As shown in

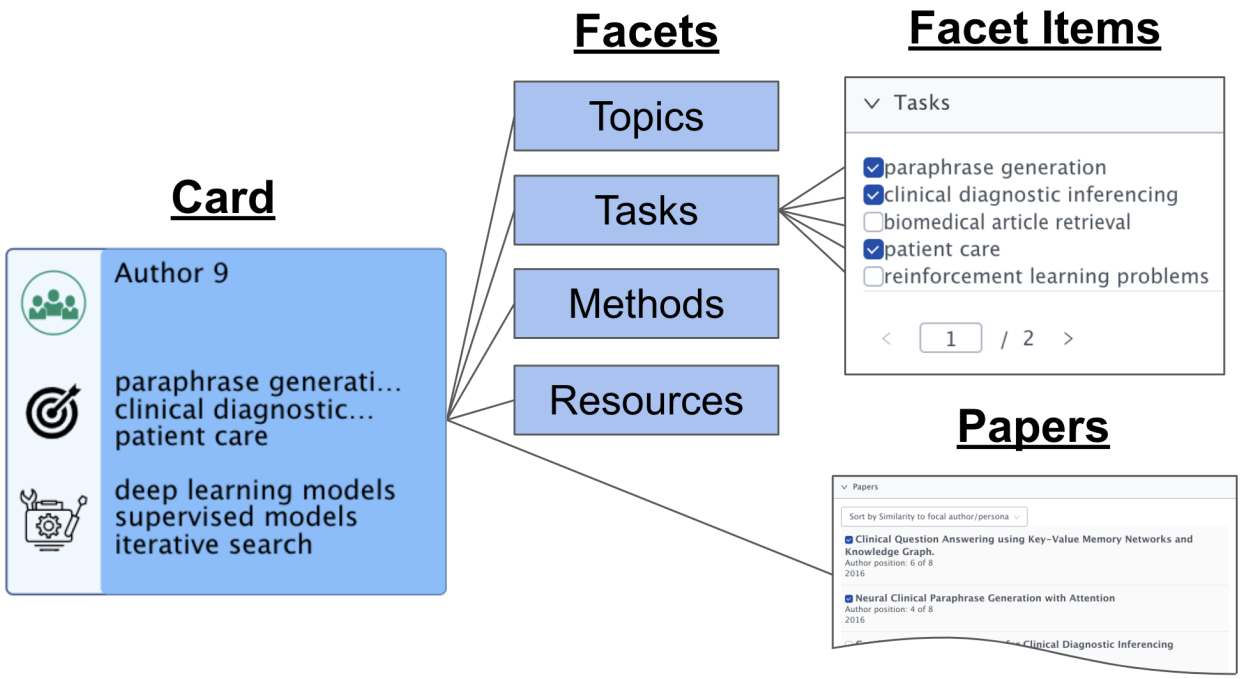

Figure 4.9: Illustration of information shown to users in Experiment II, §4.3.7. When the user clicks on an author card, an expanded view is displayed with 5 sections: papers, topics, and our extracted facets — tasks, methods, and resources.

Figure 4.9, cards showcase five sections of the author's research: their papers, MAG topics, and our extracted facet terms. We also let users view the tasks and methods ranked by *similarity* to them, which could be helpful to explain why an author was selected and better understand commonalities.

The cards showed up to five items for each section, with some sections having an optional second page, depending upon data availability. Papers could be sorted based on recency or similarity to a participant / persona. To avoid biasing participants, the only information provided for each paper was its title, date, and the suggested author's position on each paper (e.g., first, last).

Each of these items (papers and terms) had a checkbox, which the user was instructed to check if it fulfilled two criteria: 1) potentially interesting and valuable for them to learn

about or consider in terms of utility, and 2) not too similar to things they had worked on or used previously. Following a short tutorial[14], participants evaluated each author shown by checking the aforementioned checkboxes (see Figure 4.9, right). While evaluating the first and last author (randomized), the participant engaged in a protocol analysis methodology (sharing their thinking as they worked). Participants with personas were also asked, based on each persona's top five associated papers, whether they each reflected a coherent focus area, and whether they seemed useful for filtering author suggestions.

Results

For each author card evaluated by a user, we calculate the ratio of checked boxes to total boxes in that card. Then, for each user, we calculate the average of these ratios per condition (`sT`, `sTdM`, `ss`), and calculate a user-level preference S specifying which of the three conditions received the highest average ratio. Using this score, we find the proportion of users who preferred each of the `sT` and `sTdM` conditions in comparison to the `ss` approach.

Figure 4.10(a), shows results by this metric. The facet-based approaches lead to a boost over the non-faceted `ss` approach, with users overall preferring suggestions coming from the facet-based conditions. This is despite comparing against an advanced baseline geared at relevance, to which users are naturally primed.

We break down the results further by slightly modifying the metric to account for the different types of item types users could check off. In particular, we distinguish between the task/method/resource/topic checkboxes, and the paper checkboxes. For each of these two groups, we compute S in the same way, ignoring all checkboxes that are not of that type (e.g., counting only papers). This breakdown reveals a more nuanced picture. For the task, method, resource and topic facets, the gap in favor of `sT` grows considerably (Figure 4.10b). In terms of papers only, `ss`, which was trained on aggregate paper-level information, achieves a marginally better outcome compared to `sT`, with a slightly larger gap in comparison to `sTdM`

[14]All user study guidelines will be made available in our code repository.

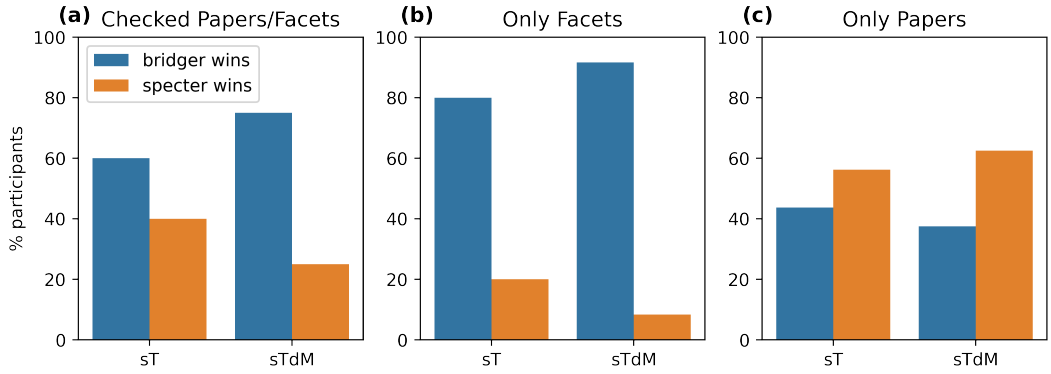

Figure 4.10: *More users prefer Bridger for suggesting novel, interesting authors.* Percent of the participants who preferred author suggestions surfaced by faceted conditions (`sT` and `sTdM`, blue bars) compared to a baseline non-faceted paper embedding (`ss`, orange bars). On average, users prefer the former suggestions, leading to more discovery of novel and valuable authors and their work (a). When broken down further, we find users substantially preferred the facet items shown for authors in our condition (b), and preferred the paper embedding baseline when evaluating papers (c). See §4.3.7 for discussion.

(Figure 4.10c). Aside from being trained on paper-level information, SPECTER also benefits from the fact that biases towards filter bubbles can be particularly strong with regard to papers. With papers, users must tease apart aspects that are new and interesting to them versus aspects that are relevant but familiar. Even if a paper is directly connected to a user's research, they may be tempted to check off a paper because they have not seen that *exact* paper or because it has minute differences from their work. In contrast, when judging a particular facet item, participants need only contemplate the novelty of the term itself, without distraction or fixation on other terms [83, 98].

Importantly, despite obtaining better results overall with the faceted approach, we stress that our goal in this paper is not to "outperform" SPECTER, but mostly to use it as a reference point — a non-faceted approach used in a real-world academic search and recommendation setting.

Item type	sT	sTdM
All	58%	75%
Paper	83%	67%
Topic	58%	75%
Task	42%	50%
Method	67%	58%
Resource	50%	67%

Table 4.1: Percentage of users with personas (N=12), for which the average proportion of checked items was higher for the persona-matched authors than for the overall-matched authors. Users saw suggested authors based on two of their personas. The suggestions came from either the `sT` or `sTdM` conditions. Reported here are counts of users who showed preference for one or both personas.

Personas We also compare the results from `sT` and `sTdM` conditions based on personas P for user $\mathcal{A}$, versus the user's non-persona-based results presented above. We compare the set of authors found using personas with authors retrieved without splitting into personas (equivalent to the union of all personas). Table 4.1 shows the number of users for which the average proportion of checked items was higher for the persona-matched authors than for the overall-matched authors (for at least one of the personas). For most participants, users signalled preference for persona-matched authors when looking at one or both of their personas. Interestingly, for papers we see a substantial boost in preference for both conditions, indicating that by focusing on more refined *slices* of the user's papers, we are able to gain better quality along this dimension too.

Evidence of Bursting Bubbles

The matched authors displayed to users were identified based either on `sT` and `sTdM` or the baseline SPECTER-based approach (`ss`). These two groups differed from each other substantially according to several empirical measures of similarity. We explore the following measures, based on author dimensions in our data that we do not use as part of the experiment: (1) Citation distance: A measure of distance in terms of citations that the user has in common

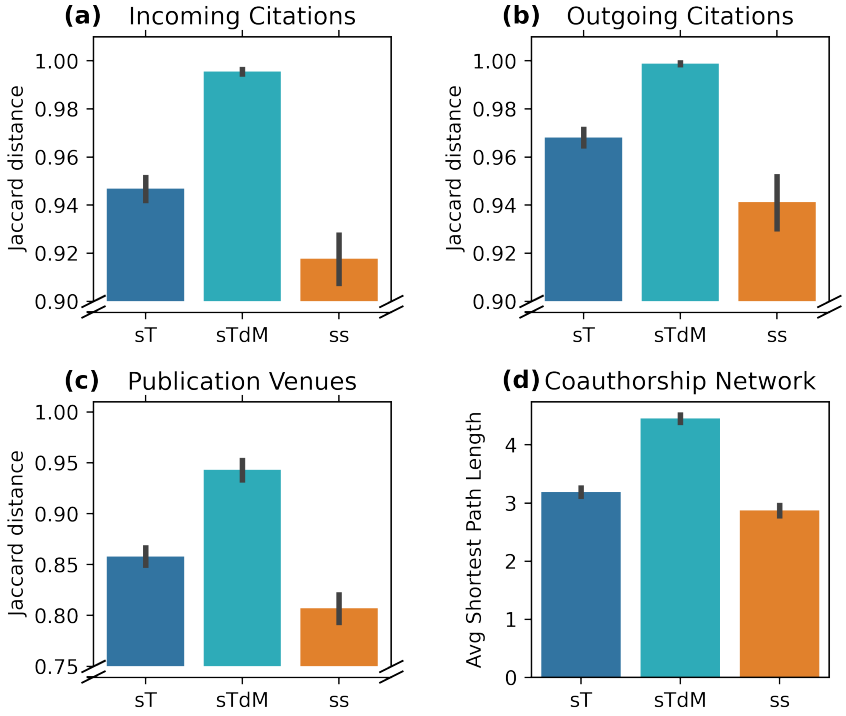

Figure 4.11: *Bridger suggests authors that are more likely to bridge gaps between communities.* In comparison to the baseline, facet-based (Bridger) author suggestions link users to broader areas. Clockwise: (a, b) Jaccard distance between suggested authors' papers and the user's papers for incoming citations (a) and outgoing citations (b); greater distance means that suggested authors are less likely to be cited by or cite the same work. (c) Jaccard distance for publication venues. (d) Shortest path length in the coauthorship graph between author and user (higher is more distant). Bridger conditions (`sT` and, especially, `sTdM`) show higher distances.

with the matched author, calculated with an intersection-over-union (Jaccard) distance. This is calculated both for incoming and outgoing citations. (2) Venue distance: The Jaccard distance between user and matched author for publication venues. (3) Coauthor shortest path: The shortest path length between the user and the matched author in the coauthorship graph. Findings of this analysis, shown in Figure 4.11, suggest that Bridger surfaces more novel authors.

In the following section, we conclude by diving deeper into user interviews we conducted, revealing more evidence and insights into user preferences and surfacing potential issues and challenges for building future author discovery systems.

4.3.8 User Interviews: Analysis & Discussion of Author Discovery

Bridges Across Scientific Filter Bubbles Bridger authors encourage more diverse connections. Under the Bridger conditions, participants formed diverse ideas that connected their research to other authors not only within their own subareas, but also other areas. As just one example, P9, who works on gradient descent for convex problems, saw a `sTdM` author's paper discussing gradient descent but for deep linear neural networks, which imply non-convex problems. They remarked, "*This is a new setup. It's very different, and it's super important.... definitely something I would like to read because it applies to things I know in a very important and hot area.*" Though the `sTdM` condition presented more of a risk in terms of surfacing authors with which the user could draw connections, it also surfaced the more far-reaching connections. P2, P6, and P19 reported similar connections. With respect to connections within the same area, participants would sometimes connect a facet of their work to the same facet of the suggested author's work. For instance, looking at a `sT` author, P15 related their research topic to another: "*I have worked on procedural text, and they follow some sequence of events, and dialogues also do that, so there is that connection.*" If participants found new connections with SPECTER, they tended to be more immediate connections to authors in their area.

Facets Help Elicit New Ideas But Require More Context Describing an author's work with short, digestible items in the form of tasks, methods, and resources helped participants find interesting new ideas. For instance, P14 expressed that a `sTdM` author's paper associated with medical image diagnosis would not be useful for them to consider because "*breaking into that space for me would require a lot of work.*" However, when they later saw 'medical image diagnosis' as a task, they commented, "*As a task, I could*

see some usefulness there. There could be other approaches that might more quickly catch my interest." Committing to interest in the task required much less effort. Moreover, participants were able to peruse more of an author's interesting tasks and methods that they did not necessarily find in their top papers. Reacting to one `sT` author, P3 did not see any papers related to 'biomedical question answering,' but they did see 'biomedical question answering system' as a method. They then noted, "*I'm going to click 'biomedical question answering' because that's not what I have worked on before, but I'm interested in learning about it.*"

Tasks, methods and resource facets support discovery better than topics. Participants were much more likely to complain that topics were too high-level to spark ideas for new, profitable research directions. P3 summarized, "*I think many of them are quite generic, so I can say I already worked on it,*" and P7 noted, "*'Artificial intelligence' is too broad. I think everything comes under that.*"

Terms with unknown meaning often garner interest, but all facets and papers require more context. Participants commonly identified tasks, methods, and resources as interesting, even when they did not fully understand their meaning. When P4 saw the method 'least-general generalization of editing examples' from a `sT` author, they stated, "*Don't know what this means exactly, but it sounds interesting.*" Nonetheless, many participants also struggled with indiscernible terms. For example, P20 said of the resource 'NAIST text corpus' under a `sT` author, "*I'm not sure what this is, and I can't guess from the name. And it wasn't mentioned in the title of the papers.*" Furthermore, some papers' titles were particularly hard to decipher. Thus, multiple participants expressed interest in having abstracts available in order to better judge their usefulness, and P15 suggested including short automated summaries [37].

Biases Toward Scientific Filter Bubbles **Time constraints in the fast-moving world of research inhibit exploration beyond the filter bubble.** Despite clear interest in an author's distant research, a couple of participants were hesitant to make connections. For example, in reacting to a `sTdM` author's paper on image segmentation for medical images,

P14 recognized, "*It'd be interesting for me to understand what is happening in that space.*" However, they added a caveat: "*I mean I'd be interested in reading [this paper] if I had infinite time.*"

Unknown background knowledge can make it intimidating to consider new areas. Engaging with distant authors' work requires a large cognitive load that can make uncovering connections difficult. P18 provided the following example: "*Maybe there's some theoretical computer science algorithm that if I knew to apply it to my problem would speed things up or something like that, but I wouldn't know enough to recognize it as interesting.*" Echoing findings in §4.3.8, this comment suggests that unfamiliar terms can especially hinder making interesting connections, and that highlighting the most useful aspects of a distant author's research may facilitate building far-reaching connections.

Preconceived notions of an area hinder consideration of connections to that area. Because Bridger's authors are selected to be more different from the user than SPECTER's authors, they often met with hard-line resistance, without full consideration of potential links. Looking at a `sTdM`-suggested author, the natural language processing (NLP) researcher P20 said in relation to three papers, "*This is not really an NLP paper, so I would pass.*" Similarly, P17 rejected `sTdM` suggestions, saying "*I don't know anything about neuroscience, and I'm not going to start now probably.*"

Personas All participants with personas stated at least one would be helpful for filtering authors. Of the 12 participants who had personas, seven described their two personas as distinct, coherent identities that would be useful for filtering author suggestions. As an example, P2 characterized their personas as related to "*human-AI collaboration or decision-making*" and "*error analysis and machine learning debugging*" respectively. Though the persona author suggestions performed relatively well in generating novel connections (Table 4.1), a few participants commented that they did not see the connection between suggested authors and their persona. For example, under a persona associated with lexical semantics, P6 commented on a `sTdM` paper, "*'Causality' is not a topic I would work on in*

lexical semantics." Diverse author suggestions may be more confusing under personas because users look for connections specific to that persona; indicating to users when these author suggestions are for exploratory purposes may be helpful.

4.3.2 Conclusion

We presented Bridger, a system and methods for facilitating discovery of novel and valuable scholars and their work. Bridger consists of a faceted author representation, allowing users to see authors who match them along certain dimensions (e.g., tasks) but not others. Bridger also provides "slices" of a user's papers, enabling them to find authors who match the user only on a subset of their papers, and only on certain facets within those papers. Our experiments with computer science researchers show that the facet-based approach was able to help users discover authors with work that is considered more interesting and novel, substantially more than a relevance-focused baseline representing state-of-art retrieval of scientific papers. Importantly, we show that authors surfaced by Bridger are indeed from more distant communities in terms of publication venues, citation links and co-authorship social ties. These results suggest a new and potentially promising avenue for mitigating the problem of isolated silos in science.

Interviews with our user-study participants show that there are many ways to improve our system. For example, we would like to improve our algorithm for persona clustering. The ability to assign informative names to personas would greatly improve usability. We also hope to address the cognitive load associated with considering new areas by providing just-in-time definitions of terms using extractive summarization [124] or generative approaches [107]. A broader challenge is in generating explanations not only for why a suggested author is found similar to the user, but also how their work may be *useful*. We also want to study whether these techniques can generalize outside of computer science, potentially connecting people with ideas from even more disparate fields as we make steps toward bridging gaps across all of science.

Chapter 5

CONCLUSION

The overwhelming scale of the scientific literature is a daunting problem. The body of existing knowledge already exceeds the capacity of any individual to keep up, and the production of new knowledge continues at an ever increasing rate. This is leading to information overload for researchers, and information silos between research groups. Despite these challenges, I am excited by the possibilities of new, technology-driven methods to help with the problem. I am heartened by the efforts of academics, industry partners like Microsoft, and nonprofits like AI2 and Meta. My contributions to this space, summarized in this disseration, are:

- An overview of scholarly data sets I have worked with, including lessons learned and code to help make use of it

- A novel method for scaling network clustering algorithms to very large citation networks

- Autoreview: a framework for building and evaluating systems to generate references for literature reviews from small sets of seed papers

- The design and evaluation of interactive, exploratory visualizations of scholarly influence over time, and of other facets of scholarly publications

- Bridger: a system for facilitating discovery of novel scholars and scientific concepts, using a novel multiview representation of authors extracted from their papers to capture their focus areas

I have been very fortunate to do this work as part of an iSchool, an institution which focuses on the intersection between people, information, and technology. Science is a large-scale coordinated human effort to further our understanding of the world. People, information, and—increasingly—technology are at the center of this endeavor. My work is firmly situated in this space, and aligned with the school's goals of connecting people with the right information, and using information to help people achieve their potential.[1] Specifically, my work with scholarly data sets is geared toward improving the access to and utility of large-scale scientific metadata so that people can make the best use of it. My work with automated literature review makes steps toward easing the burden on experts to curate the relevant papers for a topic, and also helps connect researchers with new information. My work designing exploratory visualizations tries to create opportunities for people to gain new insights from this information. And my work on SciSight / Bridger makes exciting progress toward countering filter bubbles in science, and connecting researchers in order to bridge information gaps and potentially spark new ideas.

Being part of the iSchool community has also attuned me to the potential harms that can arise from developing these technologies. Recently, numerous examples of collateral damage to people and groups have been identified as the results of algorithmically mediated systems. These examples have been brought to light through investigative journalism; they have been chronicled in books about the subject [63, 130, 129]; and a research community known as FAccT (Fairness, Accountability, and Transparency) has formed to study and address the underlying issues.[2] The impacts have been felt across many areas of modern society, including policing and sentencing [46], urban transportation [201], hiring [51], and finance [71]. These harms often arise even in cases where the technology is being explicitly developed to help—unintended consequences of biases inherent in the data used as inputs in these systems. For example, in the world of healthcare, artificially intelligent tools meant to identify and help patients with complex health needs have been shown to have significant

[1] https://ischool.uw.edu/about

[2] https://facctconference.org/

racial bias, recommending extra care to White patients over equally sick Black patients [132]. The type of work I do is not immune to these pitfalls. The tools I work on can have significant effects on the people and society it touches. Any tools which involve recommendation, for instance, such as Autoreview or Bridger, run the risk of amplifying certain content while silencing others, exacerbating rich-get-richer effects and further sidelining groups that are already marginalized. One way this could manifest would be throught the use of citation indicators as inputs, which could perpetuate gender biases in the data [193]. Tools which characterize and promote scholarly influence, such as the nautilus diagram, run similar risks. It is incumbent on researchers like myself to be vigilant to these potential harms, to audit their research and tools in order to detect them, and to take steps to address them. This is something which I have not always adequately addressed in my work. Although it is being discussed in the literature on recommendation in general [116], and although there has been some discussion around the ethics of bibliometric evaluations [26], these issues have not been much of a focus in research and development of scholarly tools. It is extremely important that I and all researchers and practitioners in this space continually strive to do better.

Information overload—the big scholarly data deluge—is the backdrop of all of this work. The creation of new scholarly information continues at an increasingly rapid pace, and does not show signs of slowing anytime soon. This reality threatens to undercut scientific progress by overwhelming researchers and obfuscating the big picture. The link between people and information weakens, and people cannot keep afloat in the sea of information. It is more important than ever to find new ways to mitigate these problems. Harnessing the scholarly literature as data has great potential for applying new technological approaches to aid scientific progress.

BIBLIOGRAPHY

[1] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. "Purpose and Polarity of Citation: Towards NLP-based Bibliometrics." In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 596–606.

[2] Wolfgang Aigner et al. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[3] Loai Albarqouni, Jenny Doust, and Paul Glasziou. "Patient preferences for cardiovascular preventive medication: a systematic review." en. In: *Heart* 103.20 (Oct. 2017), pp. 1578–1586. ISSN: 1355-6037, 1468-201X. DOI: 10.1136/heartjnl-2017-311244.

[4] Waleed Ammar et al. "Construction of the Literature Graph in Semantic Scholar." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. NAACL-HLT 2018. New Orleans - Louisiana: Association for Computational Linguistics, June 2018, pp. 84–91. DOI: 10.18653/v1/N18-3011.

[5] Douglas N. Arnold and Kristine K. Fowler. "Nefarious numbers." In: *Notices of the AMS* 58.3 (2011), pp. 434–437.

[6] Iana Atanassova, Marc Bertin, and Philipp Mayr. "Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics." In: *Frontiers in Research Metrics and Analytics* 4 (Apr. 30, 2019), p. 2. ISSN: 2504-0537. DOI: 10.3389/frma.2019.00002.

[7] Awais Athar and Simone Teufel. "Context-Enhanced Citation Sentiment Detection." In: *Proceedings of the 2012 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 597–601.

[8] Pierre Azoulay et al. "Toward a more scientific science." In: *Science* 361.6408 (Sept. 21, 2018), pp. 1194–1197. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aav2484.

[9] Jeroen Baas et al. "Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies." In: *Quantitative Science Studies* 1.1 (Feb. 2020), pp. 377–386. ISSN: 2641-3337. DOI: 10.1162/qss_a_00019.

[10] S. H. Bae et al. "Scalable Flow-Based Community Detection for Large-Scale Network Analysis." In: *2013 IEEE 13th International Conference on Data Mining Workshops*. 2013 IEEE 13th International Conference on Data Mining Workshops. Dec. 2013, pp. 303–310. DOI: 10.1109/ICDMW.2013.138.

[11] Seung-Hee Bae and Bill Howe. "GossipMap: A Distributed Community Detection Algorithm for Billion-edge Directed Graphs." In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC '15. New York, NY, USA: ACM, 2015, 27:1–27:12. ISBN: 978-1-4503-3723-6. DOI: 10.1145/2807591.2807668.

[12] Michael E. Bales, David R. Kaufman, and Stephen B. Johnson. "Evaluation of a Prototype Search and Visualization System for Exploring Scientific Communities." In: *AMIA Annual Symposium Proceedings* 2009 (2009), pp. 24–28. ISSN: 1942-597X.

[13] Michael E. Bales et al. "Bibliometric Visualization and Analysis Software: State of the Art, Workflows, and Best Practices." In: (Jan. 2020).

[14] Hilda Bastian, Paul Glasziou, and Iain Chalmers. "Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up?" en. In: *PLOS Medicine* 7.9 (Sept. 2010), e1000326. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1000326.

[15] Joeran Beel et al. "Paper recommender systems: a literature survey." In: *International Journal on Digital Libraries* 17.4 (2016), pp. 305–338.

[16] Joeran Beel et al. “Research-paper recommender systems: a literature survey.” In: *International Journal on Digital Libraries* 17.4 (Nov. 2016), pp. 305–338. ISSN: 1432-5012, 1432-1300. DOI: 10.1007/s00799-015-0156-0.

[17] Jöran Beel and Bela Gipp. “Google Scholar’s ranking algorithm: an introductory overview.” In: *Proceedings of the 12th international conference on scientometrics and informetrics (ISSI’09)*. Vol. 1. Rio de Janeiro (Brazil). 2009, pp. 230–241.

[18] Christopher W. Belter. “Citation analysis as a literature search method for systematic reviews.” en. In: *Journal of the Association for Information Science and Technology* 67.11 (2016), pp. 2766–2777. ISSN: 2330-1643. DOI: 10.1002/asi.23605.

[19] J. D. Bernal. *The Social Function of Science*. English. Main edition. London: Faber & Faber, Aug. 2010. ISBN: 978-0-571-27272-3.

[20] Vincent D Blondel et al. “Fast unfolding of communities in large networks.” In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008.

[21] Johan Bollen, Marko A. Rodriquez, and Herbert Van de Sompel. “Journal status.” In: *Scientometrics* 69.3 (Dec. 2006), pp. 669–687. ISSN: 0138-9130. DOI: 10.1007/s11192-006-0176-z.

[22] Christine L Borgman and Jonathan Furner. “Scholarly communication and bibliometrics.” In: *Annual review of information science and technology* 36.1 (2002), pp. 2–72.

[23] Katy Borner et al. “Rete-netzwerk-red: analyzing and visualizing scholarly networks using the Network Workbench Tool.” en. In: *Scientometrics* 83.3 (Jan. 2010), pp. 863–876. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-009-0149-0.

[24] Katy Börner et al. “Design and Update of a Classification System: The UCSD Map of Science.” In: *PLOS ONE* 7.7 (July 12, 2012), e39464. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0039464.

[25] Lutz Bornmann and Rüdiger Mutz. "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references." In: *Journal of the Association for Information Science and Technology* 66.11 (Nov. 2015), pp. 2215–2222. ISSN: 2330-1635. DOI: 10.1002/asi.23329.

[26] Lutz Bornmann et al. "Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results." en. In: *Ethics in Science and Environmental Politics* 8.1 (June 2008), pp. 93–102. ISSN: 1611-8014, 1863-5415. DOI: 10.3354/esep00084.

[27] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. "D3: Data-Driven Documents." In: *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2011).

[28] Kevin W. Boyack. "Using detailed maps of science to identify potential collaborations." In: *Scientometrics* 79.1 (Apr. 2009), pp. 27–44. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-009-0402-6.

[29] Jeffrey Brainard. "Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?" In: *Science* (May 13, 2020). ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abc7839.

[30] Liam Brierley. "Lessons from the influx of preprints during the early COVID-19 pandemic." English. In: *The Lancet Planetary Health* 5.3 (Mar. 2021), e115–e117. ISSN: 2542-5196. DOI: 10.1016/S2542-5196(21)00011-5.

[31] Michael Burch and Stephan Diehl. "TimeRadarTrees: Visualizing dynamic compound digraphs." In: *Computer Graphics Forum*. Vol. 27. 3. Wiley Online Library. 2008, pp. 823–830.

[32] Michael Burch et al. "Timespidertrees: A novel visual metaphor for dynamic compound graphs." In: *Visual Languages and Human-Centric Computing (VL/HCC), 2010 IEEE Symposium on*. IEEE. 2010, pp. 168–175.

[33] Ronald S. Burt. "Structural Holes and Good Ideas." In: *American Journal of Sociology* 110.2 (Sept. 2004), pp. 349–399. ISSN: 0002-9602, 1537-5390. DOI: 10.1086/421787.

[34] Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. SSRN Scholarly Paper ID 1496205. Rochester, NY: Social Science Research Network, 1992.

[35] R. K. Buter et al. "Combining concept maps and bibliometric maps: First explorations." In: *Scientometrics* 66.2 (Feb. 1, 2006), pp. 377–387. ISSN: 1588-2861. DOI: 10.1007/s11192-006-0027-y.

[36] Carter T. Butts. "Social network analysis: A methodological introduction." en. In: *Asian Journal of Social Psychology* 11.1 (Mar. 2008), pp. 13–41. ISSN: 1467-839X. DOI: 10.1111/j.1467-839X.2007.00241.x.

[37] Isabel Cachola et al. "TLDR: Extreme Summarization of Scientific Documents." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2020, pp. 4766–4777.

[38] Arie Cattan et al. "SciCo: Hierarchical Cross-Document Coreference for Scientific Concepts." In: *arXiv preprint arXiv:2104.08809* (2021).

[39] Muthu Kumar Chandrasekaran et al. "Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019)." In: *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'19. New York, NY, USA: ACM, 2019, pp. 1441–1443. ISBN: 978-1-4503-6172-9. DOI: 10.1145/3331184.3331650.

[40] Chaomei Chen. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." en. In: *Journal of the American Society for Information Science and Technology* 57.3 (Feb. 2006), pp. 359–377. ISSN: 15322882, 15322890. DOI: 10.1002/asi.20317.

[41] Chaomei Chen. "Expert Review. Science Mapping: A Systematic Review of the Literature." In: *Journal of Data and Information Science* 2.2 (2017), pp. 1–40. DOI: 10.1515/jdis-2017-0006.

[42] Chaomei Chen. "Science Mapping: A Systematic Review of the Literature." In: *Journal of Data and Information Science* 2.2 (Mar. 21, 2017), pp. 1–40. DOI: 10.1515/jdis-2017-0006.

[43] Li Chen et al. "How serendipity improves user satisfaction with recommendations? a large-scale user evaluation." In: *The World Wide Web Conference*. 2019, pp. 240–250.

[44] Tsung Teng Chen. "The development and empirical study of a literature review aiding system." en. In: *Scientometrics* 92.1 (July 2012), pp. 105–116. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-012-0728-3.

[45] Wanyu Chen et al. "Improving end-to-end sequential recommendations with intent-aware diversification." In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 175–184.

[46] Alexandra Chouldechova. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." In: *Big Data* 5.2 (June 2017), pp. 153–163. ISSN: 2167-6461. DOI: 10.1089/big.2016.0047.

[47] M.j. Cobo et al. "Science mapping software tools: Review, analysis, and cooperative study among tools." In: *Journal of the American Society for Information Science and Technology* 62.7 (July 1, 2011), pp. 1382–1402. ISSN: 1532-2890. DOI: 10.1002/asi.21525.

[48] Arman Cohan et al. "SPECTER: Document-level Representation Learning using Citation-informed Transformers." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 2270–2282. DOI: 10.18653/v1/2020.acl-main.207.

[49] Gordon V. Cormack and Maura R. Grossman. "Evaluation of Machine-learning Protocols for Technology-assisted Review in Electronic Discovery." In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '14. New York, NY, USA: ACM, 2014, pp. 153–162. ISBN: 978-1-4503-2257-7. DOI: 10.1145/2600428.2609601.

[50] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[51] Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women." en. In: *Reuters* (Oct. 2018).

[52] Emilio Delgado López-Cózar, Enrique Orduña-Malea, and Alberto Martín-Martín. "Google Scholar as a Data Source for Research Assessment." en. In: *Springer Handbook of Science and Technology Indicators*. Ed. by Wolfgang Glänzel et al. Cham: Springer International Publishing, 2019, pp. 95–127. ISBN: 978-3-030-02510-6 978-3-030-02511-3. DOI: 10.1007/978-3-030-02511-3_4.

[53] Ying Ding et al. "Content-based citation analysis: The next generation of citation analysis." en. In: *Journal of the Association for Information Science and Technology* 65.9 (2014), pp. 1820–1833. ISSN: 2330-1643. DOI: 10.1002/asi.23256.

[54] Hristo N. Djidjev, Grammati E. Pantziou, and Christos D. Zaroliagis. "Computing shortest paths and distances in planar graphs." In: *Automata, Languages and Programming*. Ed. by Javier Leach Albert, Burkhard Monien, and Mario Rodríguez Artalejo. Springer Berlin Heidelberg, 1991, pp. 327–338. ISBN: 978-3-540-47516-3.

[55] David Donoho. "50 Years of Data Science." In: *Journal of Computational and Graphical Statistics* 26.4 (Oct. 2, 2017), pp. 745–766. ISSN: 1061-8600. DOI: 10.1080/10618600.2017.1384734.

[56] Marian Dork et al. “PivotPaths: Strolling through faceted information spaces.” In: *Visualization and Computer Graphics, IEEE Transactions on* 18.12 (2012), pp. 2709–2718.

[57] Cody Dunne et al. “Rapid Understanding of Scientific Paper Collections: Integrating Statistics, Text Analytics, and Visualization.” In: *J. Am. Soc. Inf. Sci. Technol.* 63.12 (Dec. 2012), pp. 2351–2369. ISSN: 1532-2882. DOI: 10.1002/asi.22652.

[58] Nees Jan van Eck and Ludo Waltman. “CitNetExplorer: A new software tool for analyzing and visualizing citation networks.” In: *Journal of Informetrics* 8.4 (Oct. 2014), pp. 802–823. ISSN: 1751-1577. DOI: 10.1016/j.joi.2014.07.006.

[59] Nees Jan van Eck. “Methodological advances in bibliometric mapping of science.” PhD thesis. Rotterdam: Erasmus Research Institute of Management (ERIM), Erasmus University Rotterdam, 2011.

[60] Marc A. Edwards and Siddhartha Roy. “Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition.” en. In: *Environmental Engineering Science* 34.1 (Jan. 2017), pp. 51–61. ISSN: 1557-9018. DOI: 10.1089/ees.2016.0223.

[61] Holly Else. “How I scraped data from Google Scholar.” en. In: *Nature* (Apr. 2018). DOI: 10.1038/d41586-018-04190-5.

[62] Alessandro Epasto, Silvio Lattanzi, and Renato Paes Leme. “Ego-Splitting Framework: from Non-Overlapping to Overlapping Clusters.” en. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’17*. Halifax, NS, Canada: ACM Press, 2017, pp. 145–154. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098054.

[63] Virginia Eubanks. *Automating inequality: how high-tech tools profile, police, and punish the poor*. First Edition. New York, NY: St. Martin’s Press, 2017. ISBN: 978-1-250-07431-7.

[64] Michael Farrugia, Neil Hurley, and Aaron Quigley. "Exploring temporal ego networks using small multiples and tree-ring layouts." In: *Proc. ACHI* 2011 (2011), pp. 23–28.

[65] Michael Fire and Carlos Guestrin. "Over-optimization of academic publishing metrics: observing Goodhart's Law in action." In: *GigaScience* 8.giz053 (June 2019). ISSN: 2047-217X. DOI: 10.1093/gigascience/giz053.

[66] Eric A. Fong and Allen W. Wilhite. "Authorship and citation manipulation in academic research." en. In: *PLOS ONE* 12.12 (Dec. 2017), e0187394. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0187394.

[67] Santo Fortunato. "Community detection in graphs." In: *Physics Reports* 486.3 (Feb. 2010), pp. 75–174. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.11.002.

[68] Santo Fortunato and Darko Hric. "Community detection in networks: A user guide." In: *Physics Reports* 659 (Nov. 2016), pp. 1–44. ISSN: 03701573. DOI: 10.1016/j.physrep.2016.09.002. arXiv: 1608.00163.

[69] Santo Fortunato et al. "Science of science." In: *Science* 359.6379 (Mar. 2, 2018), eaao0185. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aao0185.

[70] Dieter Frey. "Recent research on selective exposure to information." In: *Advances in experimental social psychology* 19 (1986), pp. 41–80.

[71] Andreas Fuster et al. *Predictably Unequal? The Effects of Machine Learning on Credit Markets*. en. SSRN Scholarly Paper ID 3072038. Rochester, NY: Social Science Research Network, Oct. 2020. DOI: 10.2139/ssrn.3072038.

[72] Eugene Garfield. "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas." In: *Science* 122.3159 (July 15, 1955), pp. 108–111. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.122.3159.108.

[73] Yingqiang Ge et al. "Understanding echo chambers in e-commerce recommender systems." In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 2261–2270.

[74] Kosa Goucher-Lambert et al. "Adaptive Inspirational Design Stimuli: Using Design Output to Computationally Search for Stimuli That Impact Concept Generation." In: *Journal of Mechanical Design* 142.9 (2020).

[75] Trisha Greenhalgh and Richard Peacock. "Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources." en. In: *BMJ* 331.7524 (Nov. 2005), pp. 1064–1065. ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.38636.593461.68.

[76] Shashank Gupta and Vasudeva Varma. "Scientific Article Recommendation by Using Distributed Representations of Text and Graph." In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW '17 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, pp. 1267–1268. ISBN: 978-1-4503-4914-7. DOI: 10.1145/3041021.3053062.

[77] Suchin Gururangan et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." In: *ACL*. 2020. DOI: 10.18653/v1/2020.acl-main.740.

[78] Michael Gusenbauer and Neal R. Haddaway. "Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources." In: *Research Synthesis Methods* 11.2 (2020), pp. 181–217. ISSN: 1759-2887. DOI: 10.1002/jrsm.1378.

[79] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX." In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15.

[80] Anne-Wil Harzing. "Microsoft Academic (Search): a Phoenix arisen from the ashes?" en. In: *Scientometrics* 108.3 (Sept. 2016), pp. 1637–1647. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-016-2026-y.

[81] J. E. Hirsch. "An index to quantify an individual's scientific research output." en. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.46 (Nov. 2005), pp. 16569–16572. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0507655102.

[82] Daniel W. Hook, Simon J. Porter, and Christian Herzog. "Dimensions: Building Context for Search and Evaluation." English. In: *Frontiers in Research Metrics and Analytics* 3 (2018). ISSN: 2504-0537. DOI: 10.3389/frma.2018.00023.

[83] Tom Hope et al. "Accelerating Innovation Through Analogy Mining." In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. New York, NY, USA: Association for Computing Machinery, Aug. 4, 2017, pp. 235–243. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098038.

[84] Tom Hope et al. "Scaling Creative Inspiration with Fine-Grained Functional Facets of Product Ideas." In: *arXiv e-prints* (2021), arXiv–2102.

[85] Tom Hope et al. "SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search." In: *arXiv:2005.12668 [cs]* (May 26, 2020). arXiv: 2005.12668.

[86] Tanya Horsley, Orvie Dingwall, and Margaret Sampson. "Checking reference lists to find additional studies for systematic reviews." en. In: *Cochrane Database of Systematic Reviews* 8 (2011). ISSN: 1465-1858. DOI: 10.1002/14651858.MR000026.pub2.

[87] Sven E. Hug, Michael Ochsner, and Martin P. Brändle. "Citation analysis with microsoft academic." en. In: *Scientometrics* 111.1 (Apr. 2017), pp. 371–378. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-017-2247-8.

[88] Matthew Hutson. "Artificial-intelligence tools aim to tame the coronavirus literature." In: *Nature* (June 9, 2020). DOI: 10.1038/d41586-020-01733-7.

[89] A. Cecile J. W. Janssens and M. Gwinn. "Novel citation-based search method for scientific literature: application to meta-analyses." In: *BMC Medical Research Methodology* 15.1 (Oct. 2015), p. 84. ISSN: 1471-2288. DOI: 10.1186/s12874-015-0077-z.

[90] Rahul Jha, Amjad Abu-Jbara, and Dragomir Radev. "A system for summarizing scientific topics starting from keywords." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2013, pp. 572–577.

[91] Rahul Jha et al. "NLP-driven citation analysis for scientometrics." en. In: *Natural Language Engineering* 23.1 (Jan. 2017), pp. 93–130. ISSN: 1351-3249, 1469-8110. DOI: 10.1017/S1351324915000443.

[92] Marius Kaminskas and Derek Bridge. "Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems." In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7.1 (2016), pp. 1–42.

[93] Anshul Kanakia et al. "A Scalable Hybrid Research Paper Recommender System for Microsoft Academic." In: *The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: Association for Computing Machinery, May 13, 2019, pp. 2893–2899. ISBN: 978-1-4503-6674-8. DOI: 10.1145/3308558.3313700.

[94] Clint D. Kelly and Michael D. Jennions. "The h index and career assessment by numbers." In: *Trends in Ecology & Evolution* 21.4 (Apr. 2006), pp. 167–170. ISSN: 0169-5347. DOI: 10.1016/j.tree.2006.01.005.

[95] Madian Khabsa and C. Lee Giles. "The Number of Scholarly Documents on the Public Web." In: *PLOS ONE* 9.5 (May 9, 2014), e93949. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0093949.

[96] Lanu Kim, Jevin D West, and Katherine Stovel. "Echo Chambers in Science?" In: *American Sociological Association*. 2017.

[97] Lanu Kim et al. "Scientific journals still matter in the era of academic search engines and preprint archives." In: *Journal of the Association for Information Science and Technology* 71.10 (Oct. 1, 2020), pp. 1218–1226. ISSN: 2330-1635. DOI: 10.1002/asi.24326.

[98] Aniket Kittur et al. "Scaling up analogical innovation with crowds and AI." In: *Proceedings of the National Academy of Sciences* 116.6 (Feb. 5, 2019), pp. 1870–1877. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1807185116.

[99] Richard Klavans and Kevin W. Boyack. "Using global mapping to create more accurate document-level maps of research fields." In: *Journal of the American Society for Information Science and Technology* 62.1 (Jan. 1, 2011), pp. 1–18. ISSN: 1532-2882. DOI: 10.1002/asi.21444.

[100] Joel Klinger, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. "A narrowing of AI research?" In: *arXiv preprint arXiv:2009.10385* (2020).

[101] X. Kong et al. "VOPRec: Vector Representation Learning of Papers with Text Information and Structural Identity for Recommendation." In: *IEEE Transactions on Emerging Topics in Computing* (Apr. 2018), pp. 1–1. ISSN: 2168-6750. DOI: 10.1109/TETC.2018.2830698.

[102] Jean-Charles Lamirel et al. "An overview of the history of Science of Science in China based on the use of bibliographic and citation data: a new method of analysis based on clustering with feature maximization and contrast graphs." en. In: *Scientometrics* 125.3 (Dec. 2020), pp. 2971–2999. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-020-03503-8.

[103] Andrea Lancichinetti and Santo Fortunato. "Community detection algorithms: A comparative analysis." en. In: *Physical Review E* 80.5 (Nov. 2009). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.80.056117.

[104] Jill H. Larkin and Herbert A. Simon. "Why a Diagram is (Sometimes) Worth Ten Thousand Words." In: *Cognitive Science* 11.1 (Jan. 3, 1987), pp. 65–100. ISSN: 1551-6709. DOI: 10.1111/j.1551-6708.1987.tb00863.x.

[105] Kai R. Larsen et al. "Understanding the Elephant: The Discourse Approach to Boundary Identification and Corpus Construction for Theory Review Articles." In: *Journal of the Association for Information Systems* 20.7 (July 2019). ISSN: 1536-9323. DOI: 10.17705/1jais.00556.

[106] Loet Leydesdorff. "Caveats for the use of citation indicators in research and journal evaluations." en. In: *Journal of the American Society for Information Science and Technology* 59.2 (Jan. 2008), pp. 278–287. ISSN: 15322882, 15322890. DOI: 10.1002/asi.20743.

[107] Y. Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *ArXiv* abs/1907.11692 (2019).

[108] Yi Luan et al. "Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction." en. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3219–3232. DOI: 10.18653/v1/D18-1360.

[109] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. English. 1 edition. New York: Cambridge University Press, July 2008. ISBN: 978-0-521-86571-5.

[110] Alberto Martín-Martín, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. "Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison." en. In: *Scientometrics* 116.3 (Sept. 2018), pp. 2175–2188. ISSN: 1588-2861. DOI: 10.1007/s11192-018-2820-9.

[111] Justin Matejka, Tovi Grossman, and George Fitzmaurice. "Citeology: visualizing paper genealogy." In: *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2012, pp. 181–190.

[112] Philipp Mayr et al. "Introduction to the special issue on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)." In: *International Journal on Digital Libraries* 19.2 (Sept. 1, 2018), pp. 107–111. ISSN: 1432-1300. DOI: 10.1007/s00799-017-0230-x.

[113] Wes McKinney. "Data Structures for Statistical Computing in Python." In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 51–56.

[114] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks." In: *Annual review of sociology* 27.1 (2001), pp. 415–444.

[115] Rada Mihalcea and Paul Tarau. "TextRank: Bringing Order into Text." In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 404–411.

[116] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. "Recommender systems and their ethical challenges." en. In: *AI & SOCIETY* 35.4 (Dec. 2020), pp. 957–967. ISSN: 1435-5655. DOI: 10.1007/s00146-020-00950-y.

[117] Makoto Miwa et al. "Reducing systematic review workload through certainty-based screening." In: *Journal of Biomedical Informatics* 51 (Oct. 2014), pp. 242–253. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2014.06.005.

[118] Philippe Mongeon and Adèle Paul-Hus. "The journal coverage of Web of Science and Scopus: a comparative analysis." en. In: *Scientometrics* 106.1 (Jan. 2016), pp. 213–228. ISSN: 1588-2861. DOI: 10.1007/s11192-015-1765-5.

[119] James Moody, Daniel McFarland, and Skye Bender-deMoll. "Dynamic network visualization1." In: *American Journal of Sociology* 110.4 (2005), pp. 1206–1241.

[120] Tamara Munzner. "A nested model for visualization design and validation." In: *Visualization and Computer Graphics, IEEE Transactions on* 15.6 (2009), pp. 921–928.

[121] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Sept. 7, 2012. 1102 pp. ISBN: 978-0-262-30432-0.

[122] Fionn Murtagh and Pierre Legendre. "Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?" In: *Journal of classification* 31.3 (2014), pp. 274–295.

[123] Peter Mutschke et al. "Guest editors' introduction to the special issue on knowledge maps and information retrieval (KMIR)." en. In: *International Journal on Digital Libraries* 18.1 (Mar. 2017), pp. 1–3. ISSN: 1432-5012, 1432-1300. DOI: 10.1007/s00799-016-0204-4.

[124] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. "Ranking Sentences for Extractive Summarization with Reinforcement Learning." In: *NAACL-HLT*. 2018.

[125] National Academies of Sciences, Engineering, and Medicine and others. *Communicating science effectively: A research agenda*. National Academies Press, 2017.

[126] Mark Neumann et al. "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing." en. In: *undefined* (2019).

[127] Tien T Nguyen et al. "Exploring the filter bubble: the effect of using recommender systems on content diversity." In: *Proceedings of the 23rd international conference on World wide web*. 2014, pp. 677–686.

[128] Mathias Wullum Nielsen and Jens Peter Andersen. "Global citation inequality is on the rise." In: *Proceedings of the National Academy of Sciences* 118.7 (2021).

[129] Safiya Umoja Noble. *Algorithms of oppression: how search engines reinforce racism*. New York: New York University Press, 2018. ISBN: 978-1-4798-4994-9 978-1-4798-3724-3.

[130] Cathy O'Neil. *Weapons of math destruction: how big data increases inequality and threatens democracy*. First edition. New York: Crown, 2016. ISBN: 978-0-553-41881-1 978-0-553-41883-5.

[131] Alison O'Mara-Eves et al. "Using text mining for study identification in systematic reviews: a systematic review of current approaches." en. In: *Systematic Reviews* 4.1 (Dec. 2015), pp. 1–22. ISSN: 2046-4053. DOI: 10.1186/2046-4053-4-5.

[132] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations." en. In: *Science* 366.6464 (Oct. 2019), pp. 447–453. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aax2342.

[133] Lawrence Page et al. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[134] Stephen E. Palmer, ed. *Vision Science: Photons to Phenomenology*. MIT Press, 1999.

[135] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[136] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python." In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[137] Yonathan Perez et al. "Ringo: Interactive Graph Analytics on Big-Memory Machines." In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 1105–1110. ISBN: 978-1-4503-2758-9. DOI: 10.1145/2723372.2735369.

[138] J. Portenoy and J.D. West. "Dynamic Visualization of Citation Networks Showing the Influence of Scholarly Fields Over Time." In: *WWW Worskhop on Semantics, Analytics, Visualisation: Enhancing Scholarly Data*. 2016.

[139] Jason Portenoy. "Interactive visualizations and tools to explore relationships among research communities." In: *Unpublished* (May 14, 2018).

[140] Jason Portenoy, Jessica Hullman, and Jevin D. West. "Leveraging Citation Networks to Visualize Scholarly Influence Over Time." In: *Frontiers in Research Metrics and Analytics* 2 (2017). ISSN: 2504-0537. DOI: 10.3389/frma.2017.00008.

[141] Jason Portenoy and Jevin D West. "Constructing and evaluating automated literature review systems." In: *Scientometrics* 125 (2020), pp. 3233–3251.

[142] Jason Portenoy and Jevin D West. "Supervised Learning for Automated Literature Review." In: *BIRNDL 2019* (2019), p. 9.

[143] Jason Portenoy and Jevin D. West. "Constructing and evaluating automated literature review systems." In: *Scientometrics* (June 3, 2020). ISSN: 1588-2861. DOI: 10.1007/s11192-020-03490-w.

[144] Jason Portenoy and Jevin D. West. "Dynamic Visualization of Citation Networks Showing the Influence of Scholarly Fields over Time." In: *Semantics, Analytics, Visualization. Enhancing Scholarly Data*. International Workshop on Semantic, Analytics, Visualization. Springer, Cham, Apr. 11, 2016, pp. 147–151. DOI: 10.1007/978-3-319-53637-8_14.

[145] Jason Portenoy and Jevin D. West. "Visualizing Scholarly Publications and Citations to Enhance Author Profiles." In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW '17 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, pp. 1279–1282. ISBN: 978-1-4503-4914-7. DOI: 10.1145/3041021.3053058.

[146] Jason Portenoy et al. "Bridger: Toward Bursting Scientific Filter Bubbles via Novel Author Discovery." In: *In review.* 2021.

[147] Gil Press. *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. Forbes. Mar. 23, 2016. URL: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/ (visited on 04/29/2019).

[148] Jason Priem. "Beyond the paper." In: *Nature* 495.7442 (Mar. 2013), pp. 437–440. ISSN: 1476-4687. DOI: 10.1038/495437a.

[149] Helen C Purchase, Eve Hoggan, and Carsten Görg. "How important is the "mental map"?–an empirical investigation of a dynamic graph layout algorithm." In: *Graph drawing*. Springer. 2007, pp. 184–195.

[150] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: *EMNLP/IJCNLP*. 2019. DOI: 10.18653/v1/D19-1410.

[151] Karen A. Robinson et al. "Citation networks of related trials are often disconnected: implications for bidirectional citation searches." eng. In: *Journal of Clinical Epidemiology* 67.7 (July 2014), pp. 793–799. ISSN: 1878-5921. DOI: 10.1016/j.jclinepi.2013.11.015.

[152] Francesco Ronzano and Horacio Saggion. "Dr. inventor framework: Extracting structured information from scientific publications." In: *International Conference on Discovery Science*. Springer. 2015, pp. 209–220.

[153] Martin Rosvall and Carl T. Bergstrom. "Maps of random walks on complex networks reveal community structure." In: *Proceedings of the National Academy of Sciences* 105.4 (2008), pp. 1118–1123.

[154] Martin Rosvall and Carl T. Bergstrom. "Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems." en. In: *PLoS ONE* 6.4 (Apr. 2011). Ed. by Fabio Rapallo, e18209. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0018209.

[155] Seref Sagiroglu and Duygu Sinanc. "Big data: A review." In: *2013 International Conference on Collaboration Technologies and Systems (CTS)*. 2013 International Conference on Collaboration Technologies and Systems (CTS). May 2013, pp. 42–47. DOI: 10.1109/CTS.2013.6567202.

[156] Michael T. Schaub et al. "The many facets of community detection in complex networks." en. In: *Applied Network Science* 2.1 (Dec. 2017), p. 4. ISSN: 2364-8228. DOI: 10.1007/s41109-017-0023-6.

[157] Ariel S. Schwartz and Marti A. Hearst. "A simple algorithm for identifying abbreviation definitions in biomedical text." In: *Biocomputing 2003*. WORLD SCIENTIFIC, Dec. 2002, pp. 451–462. ISBN: 978-981-238-217-7. DOI: 10.1142/9789812776303_0042.

[158] Edward Segel and Jeffrey Heer. "Narrative visualization: Telling stories with data." In: *Visualization and Computer Graphics, IEEE Transactions on* 16.6 (2010), pp. 1139–1148.

[159] Filipi N. Silva et al. "Using network science and text analytics to produce surveys in a scientific topic." In: *Journal of Informetrics* 10.2 (May 2016), pp. 487–502. ISSN: 1751-1577. DOI: 10.1016/j.joi.2016.03.008.

[160] Arnab Sinha et al. "An Overview of Microsoft Academic Service (MAS) and Applications." In: ACM Press, 2015, pp. 243–246. ISBN: 978-1-4503-3473-0. DOI: 10.1145/2740908.2742839.

[161] D. J. de Solla Price. "Networks of Scientific Papers." In: *Science* 149.3683 (July 30, 1965), pp. 510–515. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.149.3683.510.

[162] Andreas Strotmann and Dangzhi Zhao. "Author name disambiguation: What difference does it make in author-based citation analysis?" en. In: *Journal of the American Society for Information Science and Technology* 63.9 (2012), pp. 1820–1833. ISSN: 1532-2890. DOI: 10.1002/asi.22695.

[163] Shivashankar Subramanian et al. "S2AND: A Benchmark and Evaluation System for Author Name Disambiguation." In: *arXiv preprint arXiv:2103.07534* (2021).

[164] Jie Tang et al. "Cross-domain collaboration recommendation." In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, pp. 1285–1293.

[165] Xuli Tang et al. "The pace of artificial intelligence innovations: Speed, talent, and trial-and-error." In: *Journal of Informetrics* 14.4 (2020), p. 101094.

[166] Guy Tsafnat et al. "The automation of systematic reviews: Would lead to best currently available evidence at the push of a button." In: *BMJ: British Medical Journal* 346.7891 (2013), pp. 8–8. ISSN: 0959-8138.

[167] Chun-Hua Tsai and Peter Brusilovsky. "Beyond the ranked list: User-driven exploration and diversification of social recommendation." In: *23rd international conference on intelligent user interfaces*. 2018, pp. 239–250.

[168] Chun-Hua Tsai et al. "Diversity Exposure in Social Recommender Systems: A Social Capital Theory Perspective." In: *work* 5.11 (2020), p. 22.

[169] Barbara Tversky, Julie Bauer Morrison Y, and Mireille Betrancourt. "Animation: Can it facilitate." In: *International Journal of Human-Computer Studies* 57 (2002), pp. 247–262.

[170] Brian Uzzi et al. "Atypical Combinations and Scientific Impact." In: *Science* 342.6157 (Oct. 25, 2013), pp. 468–472. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1240474.

[171] Marco Valenzuela, Vu A. Ha, and Oren Etzioni. "Identifying Meaningful Citations." In: *AAAI Workshop: Scholarly Big Data*. 2015.

[172] Richard Van Noorden. "Google Scholar pioneer on search engine's future." en. In: *Nature News* (Nov. 2014). DOI: 10.1038/nature.2014.16269.

[173] Corinna Vehlow, Fabian Beck, and Daniel Weiskopf. "The state of the art in visualizing group structures in graphs." In: *Eurographics Conference on Visualization (EuroVis)-STARs*. Vol. 2. The Eurographics Association, 2015.

[174] Corinna Vehlow et al. "Radial layered matrix visualization of dynamic graphs." In: *Information Visualisation (IV), 2013 17th International Conference*. IEEE. 2013, pp. 51–58.

[175] Daril Vilhena et al. "Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication." In: *Sociological Science* 1 (2014), pp. 221–238. ISSN: 23306696. DOI: 10.15195/v1.a15.

[176] Martijn Visser, Nees Jan van Eck, and Ludo Waltman. "Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic." In: *Quantitative Science Studies* (Jan. 2021), pp. 1–22. DOI: 10.1162/qss_a_00112.

[177] David Wadden et al. "Entity, Relation, and Event Extraction with Contextualized Span Representations." In: *EMNLP/IJCNLP*. 2019. DOI: 10.18653/v1/D19-1585.

[178] Caroline S. Wagner et al. "Do Nobel Laureates Create Prize-Winning Networks? An Analysis of Collaborative Research in Physiology or Medicine." In: *PLOS ONE* 10.7 (July 2015), e0134164. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0134164.

[179] Dylan Walker et al. "Ranking scientific publications using a model of network traffic." In: *Journal of Statistical Mechanics: Theory and Experiment* 2007.06 (June 2007), P06010–P06010. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2007/06/P06010.

[180] Byron C. Wallace et al. "Semi-automated screening of biomedical citations for systematic reviews." In: *BMC Bioinformatics* 11.1 (Jan. 2010), p. 55. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-55.

[181] Ludo Waltman. "A review of the literature on citation impact indicators." In: *Journal of Informetrics* 10.2 (May 2016), pp. 365–391. ISSN: 1751-1577. DOI: 10.1016/j.joi.2016.02.007.

[182] Huaiyu Wan et al. "Aminer: Search and mining of academic social networks." In: *Data Intelligence* 1.1 (2019), pp. 58–76.

[183] Kuansan Wang et al. "A Review of Microsoft Academic Services for Science of Science Studies." In: *Frontiers in Big Data* 2 (2019). ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00045.

[184] Kuansan Wang et al. "Microsoft Academic Graph: When experts are not enough." In: *Quantitative Science Studies* 1.1 (Jan. 23, 2020), pp. 396–413. DOI: 10.1162/qss_a_00021.

[185] Lucy Lu Wang et al. "CORD-19: The Covid-19 Open Research Dataset." In: *ArXiv* (2020).

[186] Ningxia Wang, Li Chen, and Yonghua Yang. "The Impacts of Item Features and User Characteristics on Users' Perceived Serendipity of Recommendations." In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 2020, pp. 266–274.

[187] Wei Wang et al. "Sustainable collaborator recommendation based on conference closure." In: *IEEE Transactions on Computational Social Systems* 6.2 (2019), pp. 311–322.

[188] Jane Webster and Richard T. Watson. "Analyzing the Past to Prepare for the Future: Writing a Literature Review." In: *MIS Quarterly* 26.2 (2002), pp. xiii–xxiii. ISSN: 0276-7783.

[189] J. D. West, I. Wesley-Smith, and C. T. Bergstrom. "A Recommendation System Based on Hierarchical Clustering of an Article-Level Citation Network." In: *IEEE Transactions on Big Data* 2.2 (June 2016), pp. 113–123. DOI: 10.1109/TBDATA.2016.2541167.

[190] Jevin D West, Ian Wesley-Smith, and Carl T Bergstrom. In: *IEEE Transactions on Big Data* 2.2 (June 2016), pp. 113–123. DOI: 10.1109/TBDATA.2016.2541167.

[191] Jevin D. West, Theodore C. Bergstrom, and Carl T. Bergstrom. "The Eigenfactor MetricsTM: A Network Approach to Assessing Scholarly Journals." en. In: *College & Research Libraries* 71.3 (May 2010), pp. 236–244. ISSN: 0010-0870, 2150-6701. DOI: 10.5860/0710236.

[192] Jevin D. West and Jason Portenoy. "Delineating Fields Using Mathematical Jargon." In: *BIRNDL 2016 Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries* 13 (2016), p. 14.

[193] Jevin D. West et al. "The Role of Gender in Scholarly Authorship." en. In: *PLOS ONE* 8.7 (July 2013), e66212. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0066212.

[194] Ryan Whalen et al. "Natural Language Processing, Article Content & Bibliometrics: Predicting High Impact Science." In: *ASCW* 15 (2015), pp. 6–8.

[195] Ryen W. White and Resa A. Roth. "Exploratory Search: Beyond the Query-Response Paradigm." In: *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1.1 (Jan. 1, 2009), pp. 1–98. ISSN: 1947-945X. DOI: 10.2200/S00174ED1V01Y200901ICR003.

[196] Mark Wilhelm et al. "Practical diversified recommendations on youtube with determinantal point processes." In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2018, pp. 2165–2173.

[197] Kyle Williams et al. "Scholarly big data information extraction and integration in the CiteSeerχ digital library." In: *2014 IEEE 30th International Conference on Data Engineering Workshops*. 2014 IEEE 30th International Conference on Data Engineering Workshops (ICDEW). Chicago, IL, USA: IEEE, Mar. 2014, pp. 68–73. ISBN: 978-1-4799-3481-2. DOI: 10.1109/ICDEW.2014.6818305.

[198] Dietmar Wolfram. "Bibliometrics, Information Retrieval and Natural Language Processing: Natural Synergies to Support Digital Library Research." In: *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*. June 2016, pp. 6–13.

[199] Zhaohui Wu et al. "Towards building a scholarly big data platform: Challenges, lessons and opportunities." In: *IEEE/ACM Joint Conference on Digital Libraries*. IEEE/ACM Joint Conference on Digital Libraries. Sept. 2014, pp. 117–126. DOI: 10.1109/JCDL.2014.6970157.

[200] Feng Xia et al. "Big Scholarly Data: A Survey." In: *IEEE Transactions on Big Data* 3.1 (Mar. 1, 2017), pp. 18–35. ISSN: 2332-7790. DOI: 10.1109/TBDATA.2016.2641460.

[201] An Yan and Bill Howe. "Fairness-Aware Demand Prediction for New Mobility." en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.01 (Apr. 2020), pp. 1079–1087. ISSN: 2374-3468. DOI: 10.1609/aaai.v34i01.5458.

[202] Erjia Yan and Ying Ding. "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other." en. In: *Journal of the American Society for Information Science and Technology* 63.7 (July 2012), pp. 1313–1326. ISSN: 1532-2890. DOI: 10.1002/asi.22680.

[203] Erjia Yan and Cassidy R. Sugimoto. "Institutional interactions: Exploring social, cognitive, and geographic relationships between institutions as demonstrated through citation networks." en. In: *Journal of the American Society for Information Science and Technology* 62.8 (Aug. 2011), pp. 1498–1514. ISSN: 1532-2890. DOI: 10.1002/asi.21556.

[204] Ka-Ping Yee et al. "Animated Exploration of Dynamic Graphs with Radial Layout." In: *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*. INFOVIS '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 43–. ISBN: 0-7695-1342-5.

[205] Zhe Yu, Nicholas A. Kraft, and Tim Menzies. "Finding better active learners for faster literature reviews." en. In: *Empirical Software Engineering* 23.6 (Dec. 2018), pp. 3161–3186. ISSN: 1573-7616. DOI: 10.1007/s10664-017-9587-0.

[206] Zhe Yu and Tim Menzies. "FAST2: An intelligent assistant for finding relevant papers." In: *Expert Systems with Applications* 120 (Apr. 2019), pp. 57–71. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.11.021.

[207] An Zeng et al. "The science of science: From the perspective of complex systems." In: *Physics Reports* 714-715 (Nov. 16, 2017), pp. 1–73. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2017.10.001.

[208] Pengfei Zhao and Dik Lun Lee. "How much novelty is relevant? it depends on your curiosity." In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2016, pp. 315–324.

[209] Xiaodan Zhu et al. "Measuring academic influence: Not all citations are equal." en. In: *Journal of the Association for Information Science and Technology* 66.2 (2015), pp. 408–427. ISSN: 2330-1643. DOI: 10.1002/asi.23179.

[210] Ziwei Zhu, Jianling Wang, and James Caverlee. "Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems." In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 449–458.

[211] Michel Zitt. "Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation." In: *Scientometrics* 102.3 (Mar. 2015), pp. 2223–2245. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-014-1482-5.