Big Scholarly Data: A Survey

Feng Xia, *Senior Member, IEEE*, Wei Wang, Teshome Megersa Bekele, Huan Liu, *Fellow, IEEE*

Abstract—With the rapid growth of digital publishing, harvesting, managing, and analyzing scholarly information have become increasingly challenging. The term Big Scholarly Data is coined for the rapidly growing scholarly data, which contains information including millions of authors, papers, citations, figures, tables, as well as scholarly networks and digital libraries. Nowadays, various scholarly data can be easily accessed and powerful data analysis technologies are being developed, which enable us to look into science itself with a new angle. In this paper, we examine the background and state of the art of big scholarly data. We first introduce the background of scholarly data management and relevant technologies. Secondly, we review data analysis methods, such as statistical analysis, social network analysis, and content analysis for dealing with big scholarly data. Finally, we look into representative research issues in this area, including scientific impact evaluation, academic recommendation, and expert finding. For each issue, the background, main challenges, and latest research are covered. These discussions aim to provide a general overview and big picture to scholars interested in this emerging area. This survey paper concludes with a discussion of open issues and promising future directions.

Index Terms—Scholarly data, Data analysis, Academic social networks, Science of science.

1 Introduction

Recent years have witnessed the rapidly growing scholarly information due to vast research works are undertaken in academia and industry [1], [2]. Research progress and results are usually articulated through publishing articles. As a result of advancement in science, scientists around the world steadily produce a large volume of research articles, which provide the technological basis for worldwide dissemination of scientific findings. In addition, researchers share their teaching materials such as slides and partial description of projects, patents and books through their homepages. The term Big Scholarly Data (BSD) is coined for this rapidly growing scholarly source of information. Large collections of scholarly data have millions of authors, papers, citations, figures, tables, etc., as well as massive scale related data such as scholarly networks, digital libraries, etc. [3], [4], [5]. BSD stands for the vast quantity of data that is associated with scholarly undertakings, such as journal articles, conference proceedings, degree theses, books, patents, presentation slides, and experimental data [6].

In the academic landscape, in general, the rapid rise of BSD brings about new issues and challenges with respect to data management and analysis. According to [7], [8], big data is characterized by four key aspects: volume, variety, velocity, and value. Based on the data properties of BSD, it is characterized by the 5V feature, where the veracity is added. The 5V feature of BSD can be seen from Fig. 1. The feature of veracity is added because data veracity is certainly important in scholarly data for issues such as author disambiguation and deduplication [9]. Using Microsoft Academic Search and Google Scholar, Williams et. al [1] estimated that there are at least 114 million English-language scholarly documents or their records accessible on The Web and also stated that new scholarly documents are generated at a rate of tens of thousands per day, which is evidence for volume and velocity of BSD [10]. The variety of BSD is drawn from the fact that it encompasses various entities and relationships among these entities, which makes the scholarly data a complex system [6], as shown in Fig. 2.

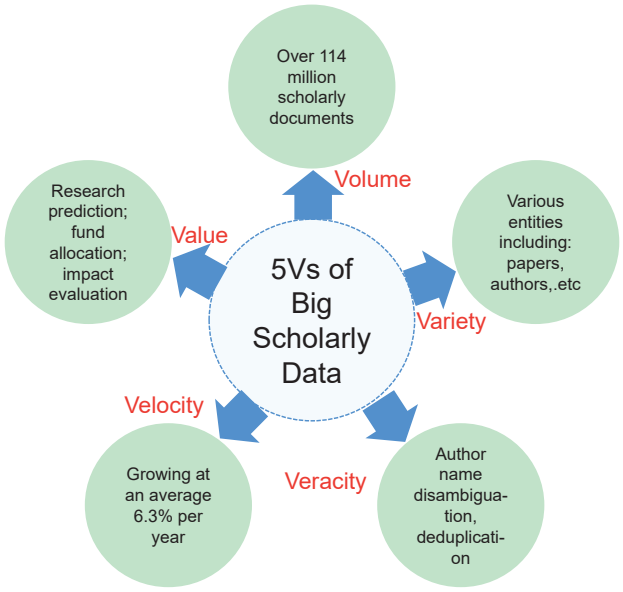


Fig. 1. The 5V feature of Big Scholarly Data

Exploring BSD can provide great benefits for various stakeholders [11], [12], [13], [14], [15], [16]. The basic motivation is to mine knowledge from BSD to provide better academic services for scholars and understand the rules and laws of science itself. For example, by analyzing the citation relationships extracted from large collections of papers, we may evaluate the impact of a given paper or scholar, which can help to allocate reputations to scientists [17]. Similarly,

F. Xia, W. Wang and T.M. Bekele are with School of Software, Dalian University of Technology, Dalian 116620, China.
H. Liu is with School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, 85281, USA.
Corresponding author: Feng Xia; e-mail: f.xia@ieee.org.

by analyzing the coauthor behaviors among scholars, we can find the distribution of scientific communities, which can help to build the map of sciences. Meanwhile, BSD analysis is not only important for academia but also helps in planning and development, i.e., for sociologists to understand researcher interactions [18], for policy makers to address knowledge and resources sharing, and for scientists, businesses, and the general public as a reference.

BSD analysis aims to solve the problems under the scope of Science of Sciences [19]. BSD analysis uses heterogeneous datasets and advanced data mining and visualization algorithms to improve our understanding of the structure and dynamics of science. This analysis utilizes 'big data', i.e., large, complex, and diverse datasets. Frequently, different types of methods have been applied to truly understand science itself which is a social system. They adopt data mining to uncover hidden patterns, correlations, and laws. Disentangling BSD analysis challenges is of great importance to promoting the development of academic society. From an analytical perspective, BSD analysis is challenging because of the increasing data volumes and various data types. The deluge of scholarly data is a direct result of advances in technology and the computerization of every aspect of modern science. Therefore, it is vital for scholars to extract concise insights from available data for specific goals. This is where the task of scholarly data mining comes in.

Even though in-depth analysis of BSD may provide an immense understanding of science and scientific communities, little research has been done on this topic. Previous tools and technologies cannot meet the requirement for dealing with BSD. Scholars cannot get access to enough scholarly information. Fortunately, with the development of big data analysis and related technologies, now we can better understand BSD and make better use of it. Meanwhile, more and more scholars, institutions, and digital libraries have made available data and codes open-source in support replication of experiments.

In this paper, we present a survey of the emerging field of BSD. To the best of our knowledge, this paper is the first effort in providing a comprehensive review of BSD. We summarize the overall research issues on BSD from three perspectives: BSD management, BSD analysis methods, and BSD applications. In the BSD management section, we review methods for scholarly data collections. Some popular digital libraries, academic search engines, and academic social networks are briefly introduced. We then discuss the methods of investigating BSD. The overall idea of exploring BSD is summarized in Fig. 3. Our aim is to provide a comprehensive understanding of research opportunities and challenges in the field of BSD and to find important issues for future explorative research on this emerging topic.

The rest of this paper is structured as follows. Section 2 presents the scholarly data management, and Section 3 describes several important data analysis methods such as statistical analysis, social network analysis, and text mining. Some promising research issues are introduced in Section 4. We discuss some critical open issues in Section 5 and conclude this paper in Section 6.

2 Big Scholarly Data Management

Given the size of scholarly information as well as its values, it is crucial to address the issues of developing scholarly focused web crawlers to harvest scholarly documents and link them to academic search engines or digital libraries. In this section, we introduce related scholarly data collection techniques, digital libraries, search engines, and academic social networks.

2.1 Scholarly Data Collection

There is evidence that large amounts of scholarly documents are freely accessible on the Web and rapidly generated daily [6]. Web data extraction systems can support the collection of these steadily growing data with minimum human effort. In documents extraction process, a web crawler should ensure that all available documents are associated with at least a valid URL where the data can be downloaded. In academic setting, scholarly documents are identified by paper metadata such as authors and venues. In Fig. 4 we depict a framework for scholarly data collection and management.

2.1.1 Scholarly Information Acquisition Framework

In Fig. 4, we depict the high-level architecture for scholarly information harvesting, indexing, and storing. The sources of scholarly documents can be the Web, homepages which contains fresh scholars' publications or profiles on academic social networks. In order to collect articles, the crawler crawls PDF files of articles from the Web. The crawler should make sure that there is, at least, a valid URL for each PDF file where the user can download before handing over to Document Extraction Component. The crawler builds a queue of documents' URLs with at least one ingestible document, schedule crawling time and update the queue to include new documents. It should also classify documents into categories such as paper and book.

Crawled documents can be automatically classified by methods such as SVM, a binary classification method that takes advantages of a paper's structural information such as document size and keywords [3]. The crawled documents then feed to Information Extractor component to mine metadata such as authors' names and venues. Once this important information is extracted, they are fed to Filtering and Categorizing component. This component performs linking metadata with articles, disambiguates author names and categorizes documents to their types such as paper and book. At the end, the framework stores and indexes the documents in the databases/repositories for users through data discovery and sharing interface.

2.1.2 Scholarly Information Extraction

Scholarly documents carry information such as metadata, citations, algorithms, figures, and tables, which are vital for developing scholarly services. Extracting this information is crucial to categorize papers into domains for the purpose of easy searching, and identifying scientific communities among others. Subsequently, we present the major scholarly information and state of the art techniques to extract them.

Metadata extraction Scholarly metadata is important to implement efficient management of scholarly documents. For extracting metadata of a paper such as title, authors,

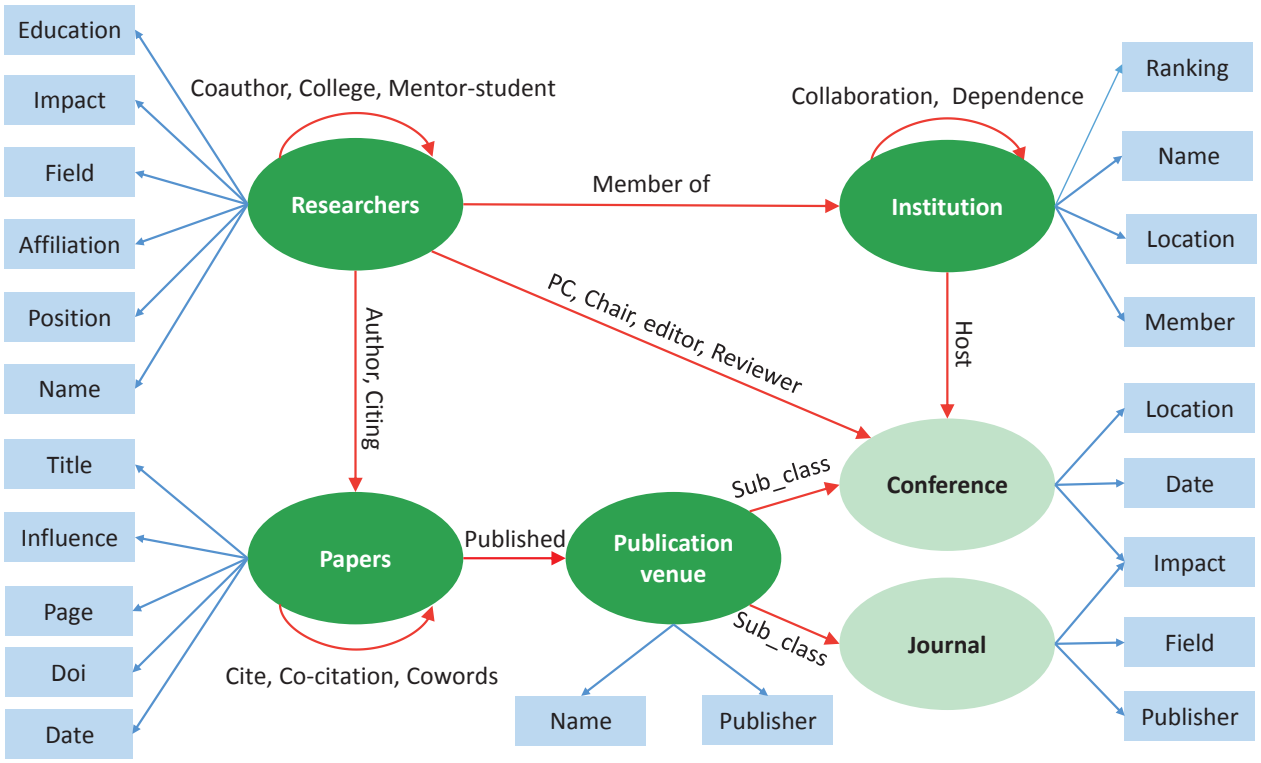


Fig. 2. Major entities and their relationships associated in Big Scholarly Data

rule-based metadata extraction is widely used. For instance, Guo and Jin [20] proposed a rule-based framework of metadata extraction where rules like header information including the title, authors and abstract are located at the first page and the title is located on the upper among others. They found that their approach worked well in finding papers but not able to handle papers with complex formats. Similarly, supervised machine learning-based techniques such as SVM or Conditional Random Fields work well with certain domains but have difficulties in dealing with multiple styles of articles. To tackle these challenges, Wu et. al [1] proposed a new mechanism with two parts: dividing and conquering multiple extractors trained for different documents types (theses, books, conferences etc.) and active learning based on web knowledge.

Citation Extraction Citations play a significant role for researchers to study the impact of scientific works, understand scientific knowledge diffusion, and identify emerging research topics [21], [22], [23]. It is also an invaluable input to build co-citation networks, to study the intellectual structure of a given domain [24]. Citation extraction requires accurately locating a section of an article with indicator 'References', 'Bibliography' or 'Sources'. These indicators can be found at the end of research articles or at the end of each chapter of a book. Once an indicator is identified, citations can be extracted parsing through the content in increasing order till the end. ParsCit is one of citation parsing tools for articles [25].

Author Detail Extraction and Profiling Contemporary digital library indexes articles and provides article searching services. An article contains a fine-grained information such as author names and affiliations. The article content is usually used to trace the research interest of the author. Author profiles can be built with article metadata constituents such as author's name, affiliations, titles, and research grants. Thus, it can help researchers to obtain a well-organized author-related information to analyze researchers interactions and scientific community detection. For instance, co-authorship networks, where nodes represent authors and edges represent papers they wrote together, can be studied to reveal how authors collaborate with each other in doing research in various contexts. However, due to journal or conference requirements, there are cases where two authors have the same name or an author uses different-form names in different venues. Hence, the appropriate linkage of articles with the right author's profile requires to solve the problem of author name ambiguity [9].

Other Information Extraction Figures [26], algorithms [27], tables, and acknowledgments are equally important to study scholarly data. Figures are usually used to present experimental results and architectures while algorithms state logical steps for experiments and system implementation. Similarly, tables describe a summary of comparing states on certain metrics or details of the data sets. The acknowledgment may have important people involved in the research but not addressed as coauthors. It also may contain information related with the fund grant for the research. These scholarly information carries unprecedented details that might be helpful for example to understand more about funding patterns of research works [28], [29].

2.2 Digital Libraries and Academic Search Engines

The rapid proliferation of information communication technology and web development technology shifts information

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2641460, IEEE Transactions on Big Data

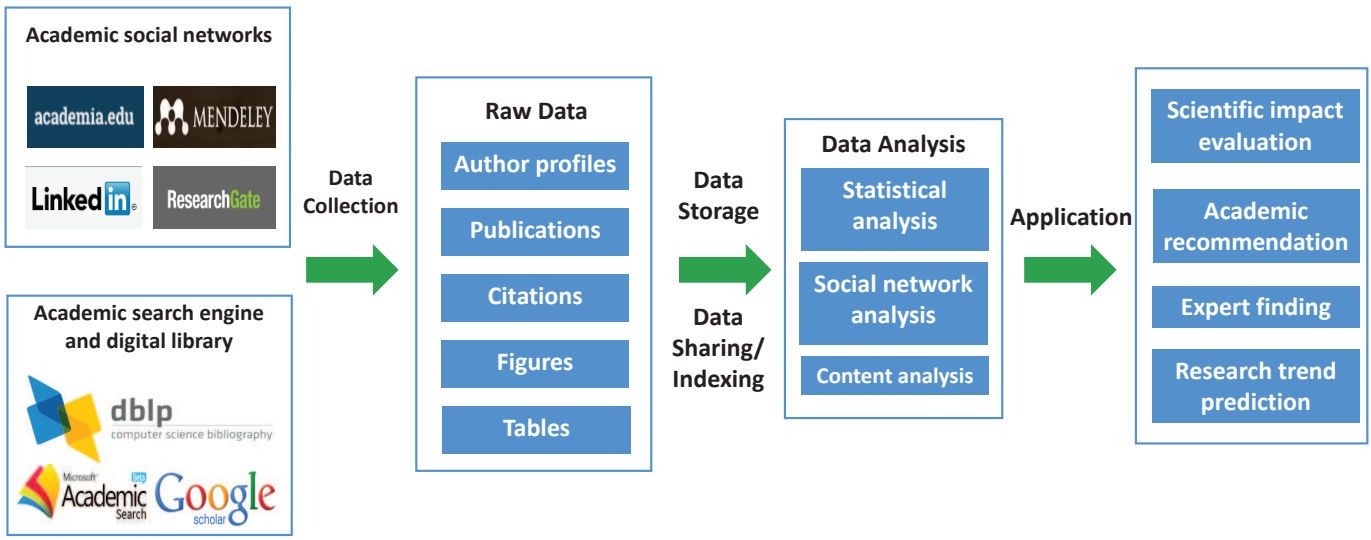


Fig. 3. Framework of Big Scholarly Data analysis

dissemination and media access to the Web. This creates an information-oriented society where a wide scope of human activities display massive information. The Web is a prevalent and interactive medium to publish, gather, and access an increasingly enormous amount of information. However, the massive information contains noises in our information accessing which impose difficulties to quick access to relevant information and affect our decisions making as well.

The explosive growth of the Web inspires librarians and information science professionals to develop reliable and effective automated systems that support an easy and effective access to the relevant information [30]. Digital libraries are such systems where the information is generated much faster than the users can process it. They are information repositories that have associated services delivered to user communities using a variety of technologies. These libraries offer different types of references and referral services, instructional services, value added services, and promotional services [30], [31].

In contemporary digital libraries, in contrast to traditional library where resources are confined within an institution in printed form, information management should integrate user interactions and heterogeneous information resources [32]. Sun and Yuan [33] described that digital libraries should serve a specific community or set of communities, conglomerate multiple entities, provide fast and efficient access with multiple access modes over time. A new metadata integration mechanism to the digital library can provide invaluable opportunities for researchers to conduct research that reveal hidden trends within these vast resources, such as research trend evolution and community dynamics. For example, CiteSeerx digital library extracts different types of metadata from scanned digital objects and PDF files on the Web and also provides services such as paper search, expert search, and collaborator recommendation as well [1], [34].

Digital libraries can play an important role in serving communities in publishing, accessing, and securing scientific information. Furthermore, they can promote the visibility and free of charge accessibility of scientific information [35]. However, with the rapid growth of scholarly data, it is challenging for users to take advantage of all the information in digital libraries. As a result, more effort has been put into the development of academic search engines that can support users to minimize the searching time and effort. Academic search engines play a crucial role to search the relevant research documents very quickly and conveniently [36]. Search engines such as Microsoft Academic Search and Google Scholar have clearly shown that search engines are enormously useful for diverse users to find research articles [37]. In Table 1, we list popular digital libraries, search engines and compare them based on some features. Besides, Semantic scholar project [1] developed by Allen Institute of Artificial Intelligence takes advantages of advanced natural language processing techniques for identifying influential citations and context extractions.

2.3 Academic Social Networks

Emerging social interaction platforms such as Facebook and Twitter play a significant role in people's daily lives. These social networks serve users as venues to share information resources and ties to others [38]. The advantages of these popular online social media and target-oriented specialized academic social network sites indicate that social networking can provide values to various types of users in different ways [39], [40]. Academic social networks can enhance sharing and disseminating of scientific knowledge and discoveries. Furthermore, they can provide platforms for scientific collaborations, promoting institution impact in education and research, and enabling scholars to share their research works and expertise.

Accordingly, after the introduction of Web 2.0, the Web provides new ways in which researchers can publish their work and communicate with each other worldwide [41]. As

1. https://www.semanticscholar.org/

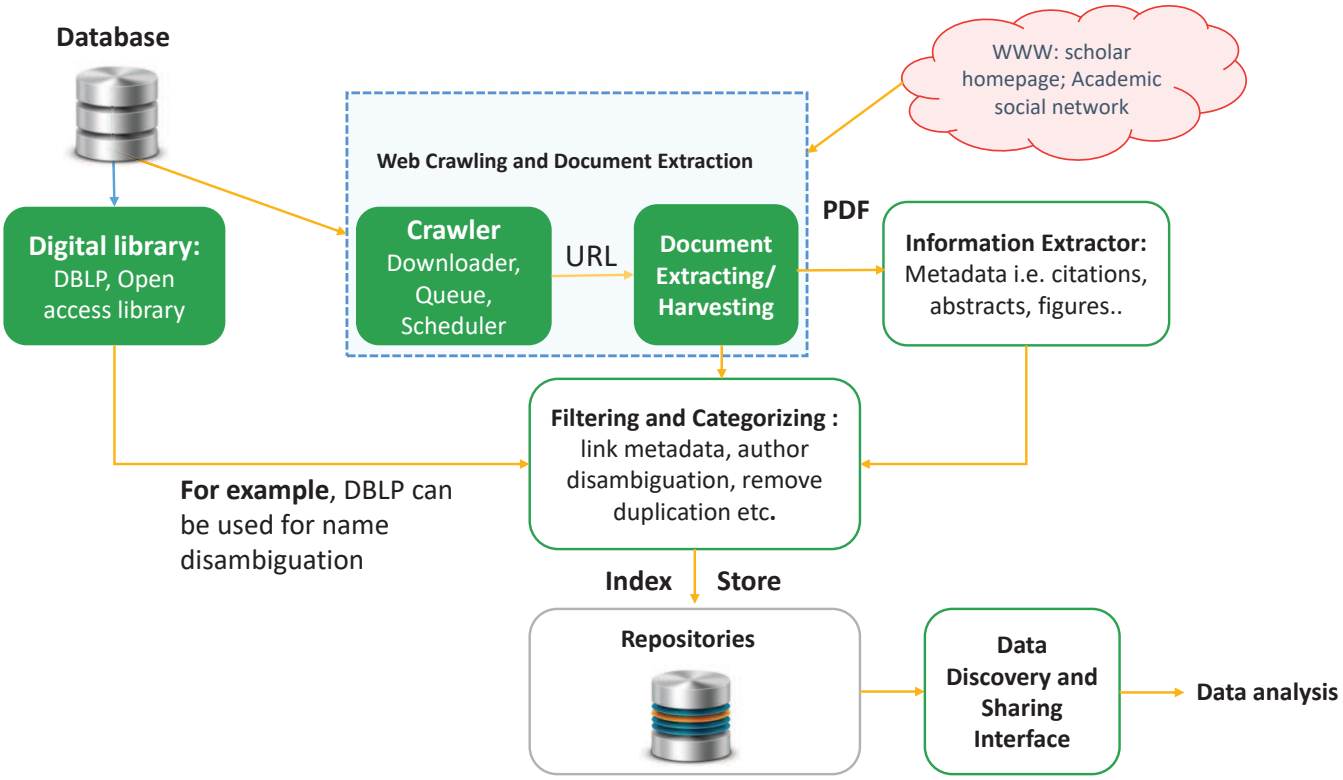


Fig. 4. High Level View of Scholarly Information Collection

authors post their articles via links on their home pages and in preprint archives, communication and connections seem to occur more naturally in email and general social network sites, such as Twitter and LinkedIn. In contrast, academic social networks such as Academia.edu and ResearchGate now combine communication and dissemination by incorporating a repository for scholarly information within a social network site for researchers [42]. Thus, they provide a new way for scholars to disseminate their publications and hence potentially change the dynamics of informal scholarly communication. In the following part of this section, we describe major academic social networks and their target audiences and community norms.

ResearchGate[2] is a social network site for researchers to create their scientific profiles, to list their publications among others and to interact with each other. It also provides researchers with a functionality to create discussion groups, share updates, results, and resources with their networks, and internal search engine that allows users to search through major databases. In addition, researchers can upload their published articles onto their personal profile pages and access events such as scientific conferences and research jobs.

Academic.edu[3] is a platform for scholars to share their research, monitor deep analytics around the impact of their research, and track the research of academics they follow in specific fields. Since its inception in September 2008, over 22 million users signed up and added about 6 million papers and 1.5 million research interests. It also attracts over 36 million unique visitors per month.

Mendeley[4] is a free reference manager and academic social network. It provides a securely storing place for users. Users can generate their citations and bibliographies in the style of their choices which are compatible with Microsoft Word, LibreOffice, and $BibTeX$. It also helps researchers to share and collaborate with each other to tackle research assignments, share feedback ,and write papers. Furthermore, researcher can connect with colleagues, peers or classmates to follow their research outputs and showcase their published research to people around the world.

VIVO[5] is an open-source interdisciplinary scientific social networking site developed by Cornell University as a platform to promote inter-disciplinary collaboration and to help recruit competitive faculty and students [41], [43]. Over time, VIVO has transformed to a platform that enables collaborations and discoveries among scientists across all disciplines. It creates a network of scientists that can facilitate scholarly discovery and allows institutions to provide semantic web-compliant data to the network.

BioWebSpin[6] is a leading professional network in Life Sciences, connecting academia with industry. It contains $100,000$ registered companies and organizations and over 10 million users. Its smart tools and boards including Dashboard, Biomatching, PubAdvanced, KOL Identification, and Job/Event Boards enable users to find and connect with the right partners, and look up information.

2. http://www.researchgate.net/
3. http://www.academia.edu/
4. https://www.mendeley.com/
5. http://www.vivoweb.org/
6. http://www.biowenspin.com/

TABLE 1
Basic information of Major Digital Libraries and Search Engines

Name	Discipline	Description	Access	Reference Management	Provider	Search Engine/Digital Library
ACM Digital Library	Computing and Information Technology	Comprehensive collection of full-text articles and bibliographic records	Subscription	No	Association for Computing Machinery	Digital library
Arnetminer	Computer Science	Comprehensive search and mining services for researcher social networks	Free	No	Tsinghua University	Both
arXiv	Multidisciplinary	Highly-automated electronic archive and distribution server for research articles	Free	No	Cornell University	Both
CiteSeerX	Computer and Information Science	Evolving scientific literature digital library and search engine	Free	No	Pennsylvania State University	Both
DBLP	Computer Science	Open bibliographic information on computer science journals and proceedings	Free	No	University of Trier	Digital library
Google Scholar	Multidisciplinary	Indexing the full text or metadata of scholarly literature across disciplines	Free	No	Google	Search engine
IEEE Xplore	Computer Science, Engineering, Electronics,	Online service used to index and search social networks	Subscription	No	IEEE Computer Society	Digital library
Mendeley	Multidisciplinary	Crowdsourced database of research documents	Free	Yes	Mendley	Search engine
Microsoft Academic Search	Multidisciplinary	Provides many innovative ways to explore scientific papers, conferences, journals, and authors.	Free	No	Microsoft Search Engines	Search engine
PubMed National	Medicine	Accessing primarily the MEDLINE database of references and abstracts on biomedical topics	Free	No	U.S. National Library of Medicine	Both
ScienceDirect	Multidisciplinary	A leading full-text scientific database offering journal articles and book chapters	Subscription	No	Elsevier	Digital library
Scopus	Multidisciplinary	A bibliographic database containing abstracts and citations for academic journal articles	Subscription	No	Elsevier	Digital library
Web of Knowledge	Multidisciplinary	An academic citation indexing and search service	Subscription	No	Thompson Reuters	Digital library

MyScienceWork[7] was created in August 2010 by Virginie Simon. Then, it has featured as a popular science media outlet dedicated to news about multidisciplinary professional research. It provides research institutes and universities with innovative platforms to share and promote scientific research. It works to make science more collaborative and allows access to a database of over 31 million scientific publications.

2.4 Data Indexing and Discovery

In many scientific disciplines, research has become increasingly data-intensive and collaborative as a result of innovations in the production and storage of large data sets. However, data sharing efforts are currently impaired by lack of proper incentives and sharing tools for data producers, practical frameworks for data standardization and indexing, and effective data discovery mechanisms [44]. As a result, currently, data generated for research analysis remain confined at their origins or are shared in a sub-optimal way just to realize the mandates of funding agencies and scientific journals. The provision of data sharing can be achieved in similar fashion with service-level agreements that define the form and quality of services in the current information technology infrastructure. As scholarly data is the most important asset for scientific research, building an uniform data-sharing agreement with incentives, policies, and tools for academic databases can enormously promote data

7. http://www.mysciencework.com/

sharing and discovery. This requires common consensus with researchers, professional societies, journal publishers, funding agencies, and information scientists to motivate users to share their data and thereby making data easily discoverable by different types of users.

In existing academic databases, Digital Object Identifiers (DOIs) are used to facilitate the data discovery and interoperability search in scholarly publications. Data Management tools require the integration of consistent and appropriate data structure to facilitate data sharing and discovery. Besides, automated tools that can index the data in line with specifications and annotations for scholarly data collections are critical for easy discovery and access to data. As data sharing inspires collaboration, journal publishers, and data repositories such as Nature, Cell, Elsevier, Springer, and PloS introduce a guideline for authors to deposit supplemental information of the data sets they used in their experiments [44]. Academic databases such as IEEE digital library and Web of Science index scholarly articles by title, abstract, key words, authors' names, conferences or journals. Meanwhile, Microsoft Academic Searches and other major web-based services provide an Application Programming Interface (API) to access data sets for research purpose.

2.5 Big Data Storage Mechanism

Most of existing academic databases manage merely conferences and journals publications. However, apart from research publications, scholarly information encompasses various scholarly outputs such as slides, books, and algorithms. These data are available in structured, semi-structured or unstructured forms. This imposes challenges in data management and analysis using traditional databases and techniques. Thus, we need to renovate academic databases and also develop a new way to collect, store, and access scholarly information. The new BSD technology should minimize hardware and processing costs and verify its value at reasonable operation resources, as well as improve performance and facilitates innovation in academic services.

Properly stored scholarly data should be accessible, safe, and manageable [45]. With the proliferation of computing technology, the enormous amount of information can be handled without requiring supercomputers and high cost. Currently, there are many management tools and techniques such as Google BigTable and Data Stream Management System (DSMS) [46]. Most widely used tools and techniques for big data are Hadoop, MapReduce, and Big Table. Detailed investigations on these tools are discussed in [47], [48]. However, academic databases and repository should investigate how to migrate and renovate their services for effectively processing large amounts of scholarly data efficiently, cost-effectively, and in a timely manner.

3 Big Scholarly Data Analysis Methods

Once we have acquired the scholarly data sets, the next step is how to analyze these data sets. The data sets collected from academic social networks, digital libraries, and academic search engines contain various entities, including text, images, graphs, and various relationships. So, how to extract useful information and detect potential scientific laws from these data sets? In this paper, we mainly consider three ways of exploring scholarly data including statistical analysis, social network analysis, and text mining technologies. In the following, we first briefly introduce four popular scholarly data sets and then introduce BSD analysis methods.

3.1 Popular Scholarly Datasets

In order to help researchers better explore academic society, many digital libraries and search engines have made available their data sets, which can be downloaded freely or by requesting access. Among these open access data sets, data sets of AMiner, Microsoft Academic Graph (MAG), DBLP, and American Physical Society (APS) are widely used for various research purposes. We have listed the basic features of these four data sets in Table 2. Among them, Aminer and DBLP mainly focus on the field of computer science. APS focuses on the field of Physics while MAG is collected from all disciplines. These data sets have been widely used to explore the science of science.

3.2 Statistical Analysis

Statistical analysis is the foundation of various methods that are used for processing datasets. It takes advantages of statistical theory. By analyzing the statistical features of data, such as mean, variance, coefficient, entropy, mathematical distribution and maximum/minimum value, scholars can find the inherent laws and regular patterns, so as to further solve critical research questions.

As a classical way of processing data, statistical analysis has been widely studied and well used nowadays. For example, Ke et al. [49] statistically analyzed the phenomenon of sleeping beauty in science. They introduce a systematic, large scale and fundamental analysis of the statistical features of sleeping beauty phenomenon. They discover the distribution of quantity of sleeping beauties, which is continuous and of power-law behavior, suggesting a common mechanism behind delayed but intense recognition at all scales.

However, data sets are becoming bigger, diversiform, and complicated. Traditional probability-based methods may not meet the demands of processing BSD. Fortunately, powerful tools and technologies have been developed including, machine learning, complex network analysis, deep learning and so on. Since the scholarly data set mainly contains two features, i.e., the network property and the text property, we will introduce these technologies from the perspectives of scholarly network analysis and scholarly text mining.

3.3 Scholarly Network Analysis

With the fast development of e-Science and Web 2.0, academic information becomes more open and easily accessed. Scientists nowadays are more dependent on scholar information than ever and various relationships among scholars have been established. An invisible social network comes into earth through academic activities, such as academic communications and collaborations, named academic social network (ASN). ASN is a special social network. How to in-depth mine the ASN effectively in the time of BSD has

TABLE 2
Basic Features of Four Popular Free-access Scholarly Data Sets

Data Set	Discipline	Size	Updated time	Downloading Link
Aminer	Computer Science	710MB	$2013-02-26$	https://aminer.org/billboard/AMinerNetwork
APS	Physics	1.21GB	$2014-07-21$	http://journals.aps.org/datasets
DBLP	Computer Science	297MB	$2015-09-05$	http://dblp.uni-trier.de/xml/
MAG	Multidisciplinary	29.8GB	$2015-08-31$	http://research.microsoft.com/en-us/projects/mag/

become an emerging topic. A potential solution is the social network analysis (SNA), which aims at studying the social relationships based on network theories. The application of SNA into ASN allows to analyze the academic relationships and to help understand the academic collaborations, as well as the citation behaviors.

3.3.1 Fundamental Network Topologies

The study of ASN involves using complex network analysis methods to investigate the topologies and dynamics of ASN, and finding out the sociological theories and laws based on network structural properties. Network topologies can be used to characterize and represent connections within a given social network. Here, we describe some fundamental network topologies that are mostly used in ASN.

Average path length: Path length is a basic metric to show the distance between two nodes in a network. Average path length can be defined as the average distance of any two nodes in a given network, which can be calculated as:

$$L = \frac{1}{\frac{1}{2}N(N-1)}\Sigma_{i\geq j}d_{ij} \quad (1)$$

where N is the number of nodes and d_{ij} is the distance between node i and node j. Based on its definition, a shorter average path length means that the network is more closely connected and information will spread faster.

Clustering coefficient: Clustering coefficient is a measure of how the nodes in a network cluster with each other. For example, in a coauthor network, with some possibilities, two friends of a given scholar may become friends with each other. Clustering coefficient can vividly depict how close your academic circle is. Clustering coefficient for a scholar i with neighbor d_i can be defined as:

$$C_i = \frac{2E_i}{d_i(d_i-1)} \quad (2)$$

where E_i is the number of edges among scholar i's neighbors.

Degree centrality: Degree centrality is a measure of the importance of a node and how influential a node is within an ASN. A node's in or out degree mean the number of connections that lead into or out of the node. Given an adjacency matrix of a graph, the degree centrality can be calculated as:

$$D_i = \Sigma_{j=1}^{n} a_{ij} \quad (3)$$

where a_{ij} is the $[i, j]$ entry of the matrix.

Many studies have been done on analyzing the statistical characteristics of ASN, such as network scope, number of publications per scholar, average number of coauthors per scholar and average number of authors per publication. At the same time, other researchers have studied the characteristics of academic social network dynamics. It has been found that as time goes on, the density of ASN becomes bigger, network diameter becomes smaller, and clustering coefficient becomes larger.

Among all the works that focus on basic statistic and topology analysis, Newman's research is the most famous one [13]. By investigating the structure of scientific collaboration networks using data drawn from a number of databases including, biomedical research, physics, and computer science, he found that collaboration networks exhibit the "small world" phenomenon. He further investigated the number of authors, mean papers per author, mean authors per paper, the number of collaborators, the giant component, average degrees of separation and clustering in scientific collaboration networks [14], [50].

3.3.2 Academic Social Network Analysis Tools

ASN analysis tools can be used to describe, analyze, and simulate an ASN by representing the characteristics of the network [51]. The main functions of ASN analysis tools include representation, visualization, characterization, and community detection of a given network. There are many tools that can be used to analyze ASN. Here, we will briefly introduce five mostly used tools namely CiteSpace, Gephi, Pajek, igraph, and NetworkX, as shown in Table 3.

CiteSpace: CiteSpace [52] is a free citation analysis tool for visualizing and analyzing trends and patterns in scientific literature. It supports structural and temporal analysis of various networks, such as collaboration networks, co-citation networks, and citation networks. The main input data source is the Web of Science. It can also be used for identifying emerging research area, finding citation hotspots, and decomposing a network into clusters.

Gephi: Gephi [53] is a widely used open source software for network analysis. It provides fruitful access to network data, such as online social networks and Email networks, and allows network clustering, spatializing, navigating and filtering. Gephi has a flexible and multi-task architecture for new possibilities to work with complex data sets and provides high-quality data visualization results.

igraph: igraph [54] is developed to handle large graphs efficiently, and can be embedded into a higher level programming language both interactively and non-interactively. It contains routines for designing, creating and visualizing networks, calculating various network properties with different file formats.

Pajek: Pajek [55], which has a long history with four versions is a widely used software for drawing networks. Pajek is a tool for analyzing large networks. It allows to handle networks with millions of nodes and edges. Pajek includes implementations for classic graph theories like

TABLE 3
Comparisons of four widely used ASNA tools

Software	Platforms	Language	Access	Features
CiteSpace	Windows/iOS	Java	Free	Visualizing and analyzing trends and patterns in scientific literature; knowledge domain visualization
Gephi	Windows/Linux/iOS	Java	Free	Exploratory Data Analysis; Social Network Analysis; Link Analysis
igraph	Windows/iOS	C/R/Python/Perl	Free	A collection of network analysis tools with the emphasis on efficiency, portability and ease of use
NetworkX	Windows/iOS	Python	Free	Creation, manipulation, and investigation of the structures, dynamics, and functions of complex networks
Pajek	Windows/iOS	C/R	Free	Analysis and visualization of large networks having some thousands or even millions of vertices

minimum spanning trees, and also implements algorithms like the community detection.

NetworkX: NetworkX [56] is a comprehensive network analysis tool. It provides the calculation of basic network features and allows integrating network structures with custom objects and data structures. Using NetworkX, standard algorithms can be used to analyze the network structure including, degree distributions, clustering coefficients, shortest paths, spectral measures, and communities.

3.3.3 Types of Scholarly Networks

Recently many studies have been done on investigating scholarly networks in data mining community. The basic motivation is to exploit knowledge from BSD to provide better academic services for scholars. From a macro sense, the static statistics and topologies of academic networks have been extensively studied. From a micro sense, extensive attentions have been paid to academic community dynamics and impact assessment of scholars. The interactions among researchers can be explored from different types of scholarly networks. Typically, there are five types of academic networks presenting academic interactions, citation networks, co-author networks, co-citation networks, co-words networks and hybrid networks [57], as can be seen from Fig. 5.

Co-author Networks Modern science is becoming more collaborative, where scholars work together across disciplines. Collaboration, presented by co-authorship, is now a ubiquitous behavior for all disciplines. Based on the co-authorship, we may construct a co-author network [58]. In co-author networks (or scientific collaboration networks), two scientists are considered connected if they have coauthored a paper. Understanding the social rules of co-author networks is especially important because it helps explore the organization of scientific communities as well as the social process of science.

According to Andrade et al. [59], the co-author networks can be classified into three categories, cross-discipline with sub-dimensions of interdisciplinary [60] and intradisciplinary [61], geographic with international [62] and intranational [63], and sector with intersector [64] and intrasector [65].

Citation Networks Another mostly investigated scientific network is the citation network, which is a kind of information network [17]. There is a basic difference between citation networks and collaboration networks because citation networks are not personal social networks, where the nodes are publications.

One of the popular goals of analyzing citation network is to measure the impact of a given paper or a scholar. Ding [66] took advantage of weighted PageRank algorithm to measure the popularity of a scholar based on a citation network. Yan et al. [67] used weighted citation to measure an article's prestige based on the assumption that weighted citations capture the popularity whereas citation counts capture the impact. Leydesdorff [68] used network centrality to measure the impact of journals.

Co-citation Networks The co-citation relationship is a phenomenon of co-occurrence in information science. If both papers A and B are cited by a paper C, they will have a co-citation relationship. And a network whose nodes have such relationships is called co-citation networks [69]. In such networks, there is a strong relationship between two linked nodes indicating that they have similar research interest or related topics.

Typically, there are mainly two types of co-citation analysis methods namely author co-citation analysis and document co-citation analysis [70]. The primary goal of co-citation network analysis is to identify the intellectual structure of a given domain [24] as well as to reveal scientific topics [23].

Bibliographic Coupling Networks Similar to co-citation network, bibliographic coupling network is also extracted from citation networks, where two papers are linked if they both cite a same article. One of the important properties of bibliographic coupling networks is that there is no delay for the calculation of the links between articles because all data needed are present upon publications.

Bibliographic coupling network has been widely used to identify research specialties, examine interdisciplinary, and map the backbone of science [57]. For example, Boyack and Klavans [71] analyzed the possibilities of using bibliographic coupling networks to detect research fronts. They further compared the accuracy of cluster solutions used for similar approaches including, co-citation analysis, bibliographic coupling, direct citation, and a bibliographic coupling-based citation-text hybrid approach.

Co-word Networks The co-word relationship is also a co-occurrence phenomenon. The network is constructed by co-words (all words or keywords), where the node represents the keywords of papers. Based on the definition, if two keywords appear in different publications at the same time, there will be a certain semantic relation among these publications as well as certain research topics among their authors [72].

In [73], Wang et al. introduced the method of building

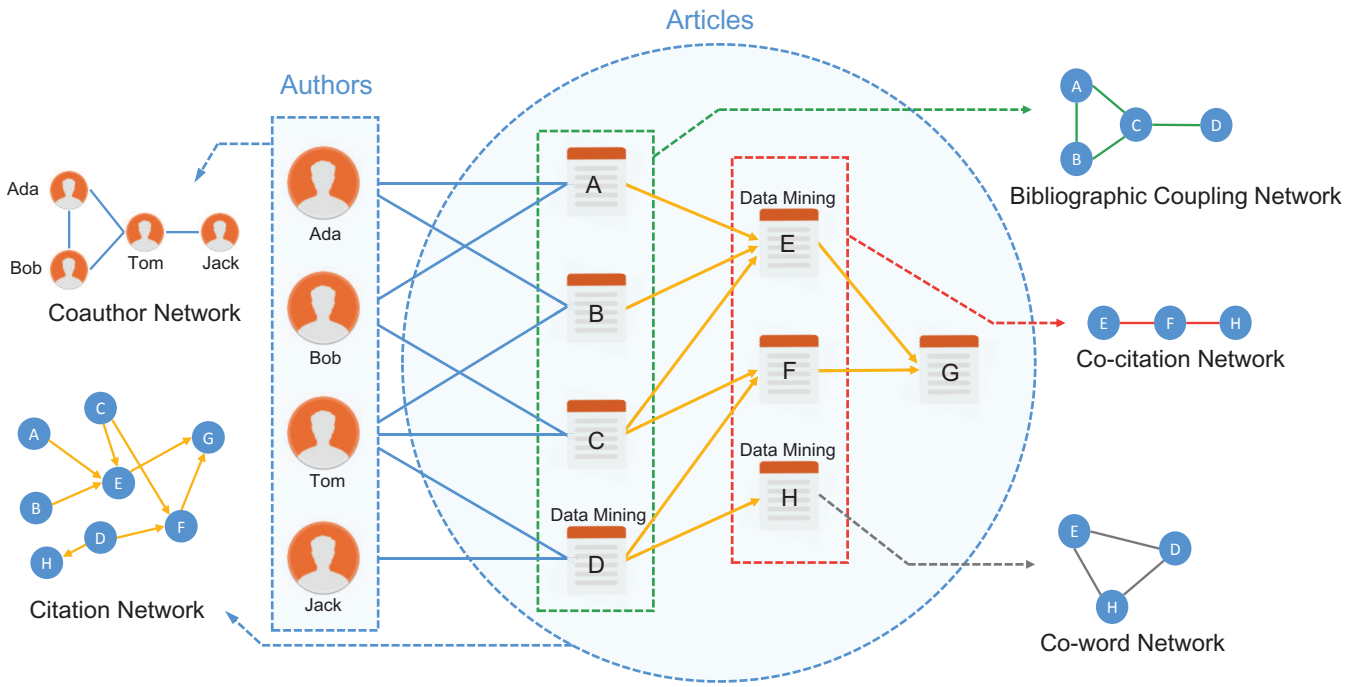


Fig. 5. Four most popular types of Scholarly networks

co-word networks. With social network analysis methods, they studied the structural properties of co-word networks, where the average distance is 2.814 and the clustering coefficient is 0.735, which demonstrates the existence of small-word characteristic.

Hybrid networks Previous four networks are homogeneous networks where the nodes are unified. However, in a real scholarly network, there are multiple types of entities (papers, authors, venues) and multiple types of relations among these entities. In this case, the nodes in a network can be papers and authors simultaneously. Networks with such characteristic are hybrid networks. Rather than focusing solely on either citation or coauthor networks hybrid networks allow us to study how people and not just papers cite in another paper.

Scholars have constructed several heterogeneous scholarly networks that can incorporate more than one entity such as bi-typed networks [74] and star-typed heterogeneous networks [75]. P-rank in [74] constructed a hybrid scholarly network containing a citation network and two co-author networks to examining influence in scholarly communication networks. Sun et al. [75] proposed a schema that contains four academic entities (papers, topics, authors, and venues) and four academic relationships (citations, publications, collaborations, and mentioning).

3.4 Scholarly Text Mining

Beside the network structure of BSD, every article is full of words, sentences and texts. Thus, mining the scholarly text data plays an important role in BSD analysis. Scholarly text mining or knowledge discovery from text focuses on the analysis of content. Since Text Data Mining is firstly introduced by Feldman and Dagan [76], the technique has been widely employed to analyze data from online social media to scientific publications [77]. The problem of text mining has gained massive attentions in recent years because of the large amounts of text data, which are created in various scholarly networks, online academic social networks and bibliographic databases. Current research in the area of text mining, as a sub-area of data mining, relies on the methods from the areas of information retrieval, information extraction, and natural language processing on order to analyze text-based corpus [78]. As a bibliographic analysis method, scholarly text mining mainly tackles the problem of large-scale topical analysis of publications covering a specific domain, institution and country [79].

3.4.1 *Textual Pattern Analysis*

Texting mining methods input raw language documents such as corpus, and output patterns, relationships, and connections related to documents. The existing methods for scholarly data mining either try to assign topics to documents based on a given keyword set (document classification) or find groups of similar documents (documents clustering). At the same time, tremendous efforts have been made to the topic-level analysis of scholarly corpus based on topic modeling algorithms.

Document classification aims at assigning pre-defined topics to text documents. For example, given a specific topic like "computer science" or "social science", document classification could automatically label each incoming publications. It mainly takes advantage of index term selection, bayes classifier, nearest neighbor classifier, decision tree classifier, and support vector machine [80]. Whatever the specific method employed, a text classification task starts with a training set $D = (d1....dn)$ of documents that are already labeled with a class $L\epsilon\lambda$ (e.g. computer science, social science). Document classification has been applied

in various domains, such as the email classification, news filtering, opinion mining, and document organization and retrieval [81].

On the other hand, document clustering takes advantage of an unsupervised learning approach to group unlabeled documents into labeled document groups, where documents within the same group are similar to one another [79]. Many approaches of document clustering are based on vector space representation, hierarchical, or partition approaches. Document clustering has been well studied. For example, Lin et al. [82] proposed a semantic document clustering method which can automatically cluster biomedical literature search result into groups for better understanding of literature search results. A more specific overview of clustering may be found in [83].

Scholarly documents contain various types of text including keywords, title, abstract, full text and so on. Through text mining approaches, we can find out the knowledge structures and scientific patterns. Analyzing the co-occurring keywords extracted from the title, abstract, or full text is one of the most widely used techniques in scholarly text mining, and has been extended to the coauthor or coheading clustering [84]. Leydesdorff et al. [85] used text mining to extract keywords from title and combined it with co-words to identify possible relationships between different contexts across different domains.

Text mining has also been applied to citation analysis to study citations from documents. Kostoff et al. [86] used this method to identify the pathways through which researcher can impact each other. Porter et al. proposed a similar research profiling approach to improve traditional literature reviews by identifying topical relationships [87]. Furthermore, researchers employed both the research profiling and journal profiling to investigate research trends [88].

Text mining has long been applied in patent analysis. Bhattacharya, Kretschmer, and Meyer [89] adopted text mining to gain co-citations and co-words between patents in order to study the connection between patents. Li et al. [90] took advantages of text mining to identify citation patterns within patents.

Apart from keyword analysis, many researchers have used text mining on full-text analysis [91], [92], [93]. Glenisson et al. combined full-text analysis and bibliometric indicators to propose a hybrid text mining method [91]. Song et al. used full-text mining to build a PubMed citation database in order to study the knowledge structure [92]. Liu et al [93] took advantage of full-text mining to identify the most significant publications given a specific domain.

3.4.2 Topical Analysis

Topic modeling has been proposed as an unsupervised method to study the contents of large document collections. The goal of topic-level analysis is to identify topics from scholarly data sets automatically by exploiting the word distribution in a corpus. The most classical model is called Latent Dirichlet Allocation (LDA) [94], which has been widely used because it provides a probabilistic model for the latent topic layer.

LDA is capable of clustering words, documents, authors, and other related entities based on latent topics. To be specific, given a document d, a multinomial distribution

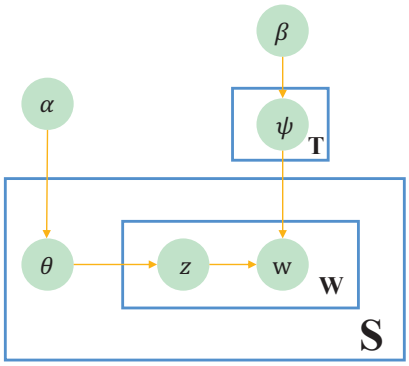


Fig. 6. The graphical description of LDA

θ_d over topics T is sampled from a Dirichlet distribution with parameter α. For each word w_{di} from document d_i, a topic t_{di} is picked from a topic multinomial distribution ψ_t sampled from a Dirichlet distribution with parameter β. Thus, we can calculate the probability of a word w from a document d as follows:

$$P(w|d,\theta,\psi) = \sum_{t \epsilon T} P(w|t,\psi_t) P(t|d,\theta_d) \tag{4}$$

Then, the likelihood of corpora C is:

$$P(T,W,|\Theta,\Psi) = \prod_{d \epsilon D} \prod_{t \epsilon T} \theta_{dt}^{n_{dt}} \times \prod_{t \epsilon T} \prod_{w \epsilon W} \psi_{tw}^{n_{tw}} \tag{5}$$

where n_{dt} is the number of times that the topic t has been mentioned in a document d, and n_{tw} represents the number of times that the word w has been associated with a topic t. In other words, the model probabilistically depicts the process of writing a paper: the scholar first chooses specific topics and then employs words that are highly related with these topics to write an article [95]. The graphical description of LDA can be seen from Fig. 6, where S denotes the whole document and z denotes a specific topic.

After the birth of LDA, it has been widely extended and used on various topics associated with scholarly data analysis [96], [97], [98]. Blei and Lafferty [96] proposed a dynamic topic model by extending classical state space models to get the topic evolution. Ding et. al [98] took advantages of topic modeling to provide topic-based article impact analysis. Tang et. al [97] applied the LDA to depict the topic distribution of authors, conferences, and citations simultaneously. They further combined the topic model with random walk framework for the academic search. Ding [98] integrated the topic modeling with path-finding to analyze the scientific collaborations and endorsement in the research area of information retrieval.

4 Big Scholarly Data Applications

In the previous section, we briefly introduce BSD analysis methods, which provide powerful approaches to processing scholarly data. At the same time, scholarly data analysis involves various applications, which can not only provide better academic services for scholars, but also help to better understand the science of science. For example, academic

recommendation systems can help scholars to overcome the information overloading problem by publication recommendations. In this section, we examine several hot research topics of BSD.

4.1 Scientific Impact Evaluation

The ability to measure scientific impact is vital for the governments and businesses which must decide how to allocate reputations and funds. Scholars also are interested in identifying the most influential papers, journals, scholars, and institutions. The measurement of scholarly impact has been experiencing rapid change with the development of scientific communications and the possibility of accessing BSD. BSD analysis has provided tools and techniques to assess scientific impact in new ways [99], [100], [101]. In this section, we highlight the scientific impact evaluation of three scientific entities including, paper, scholar, and journal.

4.1.1 Article Impact Evaluation

Evaluating the impact of a single scientific article has been extensively studied for a long history in bibliometrics and scientometrics, which helps researchers find high-quality related works. Traditional ranking methods mostly leverage citation counts [4], [102] as the base for evaluating how important an article is. However, merely citation-based ranking methods can not capture the dynamic nature of scholarly communications. Thus, many efforts have been made to employ additional information. Walker et al. [103] proposed CiteRank, which integrates the publication time into random walk model to predict future citation for each article. This model may capture the dynamics of publications by giving a high score to recent published articles. Nevertheless, this method merely used publication time and citation, which can not fully represent the impact of an article.

To tackle this problem, Sayyadi and Getoor [104] introduced a ranking model named FutureRank, which calculates the future rank score of each article by utilizing citations, authors, and time collaboratively. In FutureRank model, a new publication is expected to have a higher impact if its authors had published prestigious papers before. Furthermore, P-Rank [74] constructs a heterogenous scholarly network with various entities including, publications, authors, and journals to measure the article impact. Wang et al. [11] developed a method that ranks scientific articles by exploiting citations, authors, journals, and time information. This method employs a $PageRank + HITS$ framework to exploit different kinds of information simultaneously in heterogenous networks.

4.1.2 Author Impact Evaluation

Evaluating the cumulative impact of a scholar's research outputs is of great importance because of the limited resources in academia. Such quantification provides a reference for policy makers in university faculty recruitment and credit allocation. The publication records and citation records are obviously helpful. Hirsch [105] took the lead in quantifying author impact by proposing h-index, which is defined as the number of papers with citation number $\geq h$. Later on, g-index [106] was proposed to measure the global performance of authors as an improvement of the h-index.

Recent works begin to rank authors in heterogeneous networks with various types of nodes and relationships. A co-Ranking method [107] was proposed by Zhou et al. to rank authors and their publications between the authorship and citation networks. Meng et al. [108] introduced a Co-Rank framework which ranks authors and their publications iteratively and leverages the output of each round to reinforce the ranking of authors and papers. Furthermore, Tri-Rank was proposed by Liu et al. [109] which considers venue information to co-rank authors, papers and venues simultaneously.

4.1.3 Journal Impact Evaluation

Another important entity in scientific impact evaluation is the journal. Impact of journals in a given research can be computed using journal impact factor (JIF) [12], which is defined as the average number of citations in a year given to those articles in a journal published in the previous 2 years. However, JIF computes a mean over a heavy-tail distribution of citation counts, which may suffer from the limitation of identifying odd distribution of citations [110]. The EigenFactor metric [111] was proposed to rank journals based on PageRank on journal-journal citation graphs generated through paper citation networks.

4.2 Academic Recommendation

With the development of online academic services, such as advanced academic digital libraries, online academic social networks, and academic search engines, scholars can get access to scholarly information more easily. However, problems emerge in connection with information overloading. For example, researchers have to be well aware of recent developments in the topics they are working on. With the number of publications getting larger and larger, scholars need to choose related papers from massive potential candidates, which is time-consuming and tedious. To tackle this problem, the recommendation technology has been developed for scholars, including the academic paper (or literature) recommendation, collaboration recommendation, and venue recommendation.

4.2.1 Literature Recommendation

The academic community has produced millions of publications and the number of publications is growing all the time. Scholars have to search papers for reading and citing in order to better understand a new research topic. Literature plays a critical role in academic research. However, under the BSD context, traditional keyword-based searching method can not satisfy the requirement for most related papers. Many studies have been done to solve this problem by coming up with academic paper recommendation methods [112].

An early example of paper recommendation is proposed by McNee et al. [15]. The authors proposed a collaborative filtering-based approach for paper recommendations. They employed the citation web between papers to create the rating matrix, which is the basic component of collaborative filtering algorithm. They also investigated six algorithms for

selecting citations. Because of the advantages of collaborative filtering, their method can better solve the cold-start problem. Torres et al. [16] expanded this method by combing the collaborative filtering and content-based filtering. Their results indicate that by combining these two methods, we can improve the accuracy of the paper recommendation.

Besides the content and citation relationships, He et al. [113] developed a novel paper recommendation approach based on the citation context. They believe that a high quality paper recommendation should match the local contexts of the citations. Based on this idea, they proposed a non-parametric probability model to measure the context-relevance between a citation context and a document. By performing experiments on CiteSeerX, they find out that their methods can recommend citations for a specific context effectively.

Random Walk algorithm is applied into paper recommendation by Gori et al. [114]. They proposed a PaperRank algorithm based on the citation network and random-walker properties. Since the number of citations among papers is relatively small, the constructed citation networks are often sparse. Kucuktunc et al. [115] proposed a fast paper recommendation algorithm utilizing a sparse matrix generated from the citation graph.

At the same time, some researchers begin to use social-aware information to improve the performance of paper recommendations [116], [117]. Motivated by the importance of social characteristics of scholars in an academic social network, Xia et al. [117] proposed a folksonomy-based scholarly paper recommendation algorithm. Asabere et al. [116] further proposed a socially aware recommendation algorithm for scholarly paper recommendations.

4.2.2 Collaboration Recommendation

In academia, scientific achievements may not be reached without the collaboration among scholars. Previous Research has shown that being cooperative is a necessary characteristic for a successful researcher and researchers are becoming more cooperative [118], [119]. Therefore, it would be instrumental for scholars to get acquainted with other scholars. The academic collaborator or collaboration recommendation technology can help scholars find related scholars for collaborations.

The recommendation of academic collaborations is a special recommendation problem where two scholars are recommended to do research together [120]. For doing such recommendations, it is necessary to consider the academic relationships among researchers. For example, in [121], the authors defined two metrics to measure the relationships among researchers. The two metrics are Global Cooperation and Global Correlation. Global Cooperation is used to measure how frequently two scholars have collaborated and Global Correlation is used to measure how similar the areas of the scholars are. Based on the DBLP digital library, they construct the scientific collaboration network and evaluate their methods.

Previous studies usually formalize the academic collaboration recommendation as a link prediction problem [122], [123], [124]. The basic idea is that based on an academic social network, how to predict when will two nodes, which are not connected, connect with each other. In [122], Brandao et al. used concepts from social network analysis for collaboration recommendation in academic social networks. They proposed two new metrics considering the social principles including homophily and proximity. They focused on analyzing how these two metrics influence the recommendation performance. At the same time, Xia et al. [123] considered how to recommend most related collaborators for scholars. They proposed a novel algorithm named MVCWalker based on random walk with restart. They exploited three academic factors, i.e., coauthor order, latest collaboration time and times of collaboration when calculating the link importance between researchers. Through extensive experiments on DBLP dataset, they found that incorporating the above academic factors can improve the precision, recall rate, and the coverage rate of academic collaboration recommendations.

Interdisciplinary collaborations have become more and more popular and necessary in academic society. However, it is more difficult to establish cross-domain cooperations for researchers. Cross-domain collaboration recommendations are more challenging compared with traditional collaboration in the same domain because of the sparse connections between different research domains. Tang et al. [125] developed the Cross-domain Topic Learning (CTL) to address this problem. CTL model consolidates the cross-domain recommendation through topic layers, which can alleviate the sparseness issue. Guo and Chen [126] further studied the cross-domain collaboration recommendations by combing co-author relationships and co-citation relationships to construct networks. The experiments show that citation information can help improve the performance of cross-domain collaboration recommendations.

4.2.3 Venue Recommendation

Academic conferences do not just serve to present research progress, but also to bring scholars in the same domain together, which can foster potential collaborations. However, choosing the most related venues to attend may be time-consuming at a large conference with several parallel workshops. At the same time, scholars attending the conference are moving around, joining different talks at different rooms. Thus, how to recommend suitable venues for scholars becomes a critical problem.

In order to recommend presentation session venues at conferences, Pham et al. [127] proposed the context-aware mobile recommendation system. They combined the social context gained from academic social networks with spatio-temporal of scholars and gave venue recommendations through mobile devices. The basis of their algorithms is collaborative filtering.

Hornick et al. [128] proposed a social information recommendation system that helps scholars find out talks they may wish to listen during large academic conferences. Furthermore, Xia et al. [129] designed a socially aware venue recommendation algorithm which considers both the location and time contextual data. Their recommendation technology hybridized the computation of similar interpersonal relationships and personality traits among scholars. They used a combination of pearson correlation, social ties, contextual information, and degree centrality to generate social-aware venue recommendation for scholars. Further more, they enhanced their methods through integrating the

current context of both the smart conference community and participants in [130].

4.3 Expert Searching

Recent research trends have shown that expert searching/finding (ES) as a research issue has been given enormous attention from organizations and academia. The purpose of expert searching with a proven expertise for a given keyword depends on different contexts. Primarily, ES idea began in organizations where building knowledge base encompasses descriptions of people's skills [131] and later on is widely studied in different contexts. Following the introduction of Text Retrieval Conference (TREC) enterprise track in 2005, various works are dedicated to expert searching [132], [133], [134].

Identifying expert based on the query in associated documents requires constructing communication graphs which show the flow of information and knowledge. For example, a communication graph can be constructed between authors and articles to evaluate author's expertise in a given domain. Thus, constructing the communication graph based on the links of topic embedded in the documents is an important step. HITS, PageRank and Affinity are some of the widely used algorithms which may calculate expertise scores in the graph with/without random-walk based approaches. Based on previous research on ES, the predominantly used techniques can be categorized as: 1) profile-centric methods where an expert knowledge is directly derived from associated documents; 2) document-centric methods where first the documents are identified as per the query and then followed by locating the associated experts [131]. In the subsequent section, we describe each approach with related research works.

4.3.1 Profile-centric Method

In this approach, an expert knowledge is directly derived from associated documents. Profile-centric methods construct an expert profile as a mock document based on descriptions relevant to the expert, for example, job descriptions [131]. In academia, there is a common consensus that productive scholars are most likely considered experts. Thus, expert profile can be constructed with authors academic performance associated features such as the average publications per annum and the number of publications in journals with or without the query topics in their contents. For a given query, the ES algorithms try to find experts by matching the query with expert profiles and return a list of the most relevant experts in the order of their relevance scores [132].

4.3.2 Document-centric Methods

Most computerized ES techniques depend on document-based relevance to predict the expertise level of experts for a given domain [135], [136], [137]. This technique assumes that scholars' papers are positively related to their expertise on the query level. In contrast to profile-centric ES, in document-based ES, first relevant documents should be identified and categorized to domains prior to actually link documents to experts. Classifying research publications to domains can make expert finding easier. Keyword retrieval and unsupervised clustering are some of commonly used methods for document classification purposes. For example, researchers in [138] develop a recommendation system which utilizes both ranking and clustering methods.

In academia where researchers usually publish their findings in conferences or journals, document-centric approach is more appropriate and powerful to find experts. Taking an old notion that one's publications represent his/her expertise [50] is fundamental to search experts in bibliographic data. This data contains related information that reveals researcher's research area, quality of works (from venues where he/she published works), his/her collaborations with other researchers and fund securing history among others. Accordingly, many research works are devoted to find experts in academia based on bibliographic data [135], [139], [140].

However, the increasing research works with rapid generations of research publications and the increasing popularity of academic social networks challenge the existing approaches and algorithms to tackle the problem of locating experts in the area of BSD. As a result, researchers need to investigate new ways to address these issues through new approaches or enhancing existing approaches for ES in bibliographic data.

5 Open Issues and Challenges

In previous sections, we have surveyed several key issues associated with BSD mining including, academic recommendations, scientific impact evaluations, and the expert finding. Besides these research topics demonstrated above, there are still many open issues which are representative of critical directions both at the theoretical and the applied levels. We give a non-exhaustive, subjective lists of such issues that seem particularly promising for further research in this section.

5.1 Standard Evaluation Method

Various digital libraries and academic search engines have provided various services with different methods. For example, in order to evaluate the impact of a given scholar, a lot of ranking methods have been proposed, such as the citation, H-index, g-index, and i10-index [141]. However, different evaluation methods may have great differences. While a lot of methods of processing BSD exist, we have few ways to evaluate them. We need to develop standard rubrics, standard data sets, and benchmarks for evaluating these different methods.

5.2 Big Scholarly Data Platform

To enable the easy acquisition of sufficient BSD, academic search engines usually need to crawl useful information from the Web such as scholars' homepages and then store and index collected data. Previous client/server architecture might be able to process the data through single pipeline data processing and static crawling strategies. However, since scholarly data is growing fast, traditional systems cannot meet the demand of the high data throughput. Thus, more sophisticated scholarly data platforms other than just traditional user-oriented services should be designed to enable more advanced and useful scholarly applications.

5.3 Beyond the Publication

Previous studies of scholarly data sets mainly focus on the process of scholars writing an article. Citation relationships and coauthor relationships extracted from publications are two widely and deeply investigated directions. However, in the meantime, various other relationships have apparently not been investigated. For example, as can be seen from Fig. 2, beside coauthoring with others, scholars may be editors or reviewers in a specific conference, or be members of an institution. These positions or reputations may reflect the influence or scientific output compared with merely citation-based methods. Thus, how to gain and integrate scholars' multiple properties and relationships is a promising research topic, which may help to analyze our academic society more comprehensively.

5.4 Altmetrics

With the easy access to BSD, we can now evaluate the impact of a publication more efficiently and effectively from various aspects. We now can not only use citation to evaluate the scientific impact, but also use some other information from online social media or scholarly products, including times of commenting, downloading, and sharing, which can be defined as altmetrics [142]. However, the use of altmetrics in scientific output evaluation is still an open issue. Does social media sharing correlate with subsequent citation rates for a given article? There is a critical demand for analyzing the correlations between citations and altmetrics.

5.5 Conflict of Interest

Although the recent citation-based scientific impact evaluating methods have obtained remarkable successes, these methods may conceal anomalous citations. There may exist potential conflicts of interest (COI) relationships between scholars in citing. To be specific, COI indicates scholars or institutions involved in the same interest of various aspects, and they may deliberately cite themselves or other people with close relationships. When evaluating the scientific impact, we need to identify and analyze the COI relationships for fairness. How to define and quantify the potential COI between scholars is important and challenging.

5.6 Heterogeneous Networks Analysis

Most real-world scholarly networks are heterogeneous, containing entities of different types, such as authors, papers, venues, year of publication, and terms in a bibliographic network. Modeling co-evolution of multi-typed objects can capture richer information than that on single-typed entity alone. For example, studying the co-evolution of authors, venues, and terms in a bibliographic network can better explain the evolution of research areas than just examining co-author networks or term networks alone. Although heterogenous networks provide a richer semantic view of the data, the added complexity makes it difficult to directly apply existing techniques that work well on homogeneous networks. To further understand scientific interaction patterns and their impacts, future hot research topics and interdisciplinary research evolutions, conducting research tailored towards heterogeneous academic networks analysis is promising. The possibly research ideas are devising new methods, approaches, and techniques to bridge gaps in existing methods to analyze homogeneous networks.

6 Conclusion

BSD analysis has been accelerating in recent years. Many researchers have realized the importance of using technologies from data mining to understand scholarly data. The availability of unprecedented amounts of BSD on scientists' collaborations, documents sharing and publications open the possibility of investigating science itself as well as scientists ourselves. BSD can greatly accelerate the development of science by promoting scientific collaborations, scholar data sharing, and fair fund allocation methods. Although it is of great value to mine and analyze scholarly data, more investigations are needed to comprehensively study this topic.

In view of this, we introduce the emerging area of BSD in this survey work. We now have a good opportunity as scholars to understand and benefit academia under the BSD environments. In academia, BSD analysis is enabling researchers to do research conductively, institutions and governments to move away from experience-based to data-driven policy design. It is time to take advantage of the power of BSD to promote the development of novel learning technologies to advance science and technology.

References

[1] Z. Wu, J. Wu, M. Khabsa, K. Williams, H.-H. Chen, W. Huang, S. Tuarob, S. R. Choudhury, A. Ororbia, P. Mitra *et al.*, "Towards building a scholarly big data platform: Challenges, lessons and opportunities," in *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*. IEEE, 2014, pp. 117–126.

[2] J. Priem, "Scholarship: Beyond the paper," *Nature*, vol. 495, no. 7442, pp. 437–440, 2013.

[3] C. Caragea, J. Wu, K. Williams, S. Das, M. Khabsa, P. Teregowda, and C. L. Giles, "Automatic identification of research articles from crawled documents," *Proceedings of WSDM-WSCBD*, 2014.

[4] S. Lehmann, A. D. Jackson, and B. E. Lautrup, "Measures for measures," *Nature*, vol. 444, no. 7122, pp. 1003–1004, 2006.

[5] Y.-R. Lin, H. Tong, J. Tang, and K. S. Candan, "Guest editorial: Big scholar data discovery and collaboration," *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 1–2, 2016.

[6] K. Williams, J. Wu, S. R. Choudhury, M. Khabsa, and C. L. Giles, "Scholarly big data information extraction and integration in the citeseer χ digital library," in *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*. IEEE, 2014, pp. 68–73.

[7] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. IEEE, 2013, pp. 995–1004.

[8] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*. IEEE, 2013, pp. 42–47.

[9] A. A. Ferreira, M. A. Gonçalves, and A. H. Laender, "A brief survey of automatic methods for author name disambiguation," *Acm Sigmod Record*, vol. 41, no. 2, pp. 15–26, 2012.

[10] M. Khabsa and C. L. Giles, "The number of scholarly documents on the public web," *PloS one*, vol. 9, no. 5, p. e93949, 2014.

[11] Y. Wang, Y. Tong, and M. Zeng, "Ranking scientific articles by exploiting citations, authors, journals, and time information," in *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[12] E. Garfield, "The history and meaning of the journal impact factor," *Jama*, vol. 295, no. 1, pp. 90–93, 2006.

[13] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[14] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[15] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl, "On the recommending of citations for research papers," in *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 2002, pp. 116–125.

[16] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl, "Enhancing digital libraries with techlens+," in *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2004, pp. 228–236.

[17] M. Newman, *Networks: an introduction*. Oxford University Press, 2010.

[18] W. Wang, J. Liu, S. Yu, C. Zhang, Z. Xu, and F. Xia, "Mining advisor-advisee relationships in scholarly big data: A deep learning approach," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, 2016, pp. 209–210.

[19] R. P. Light, D. E. Polley, and K. Börner, "Open data and open code for big science of science studies," *Scientometrics*, vol. 101, no. 2, pp. 1535–1551, 2014.

[20] Z. Guo and H. Jin, "A rule-based framework of metadata extraction from scientific papers," in *Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2011 Tenth International Symposium on*. IEEE, 2011, pp. 400–404.

[21] J. Huang, Z. Zhuang, J. Li, and C. L. Giles, "Collaboration over time: characterizing and modeling network evolution," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 2008, pp. 107–116.

[22] Z. Shen, K.-L. Ma, and T. Eliassi-Rad, "Visual analysis of large heterogeneous social networks by semantic and structural abstraction," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 12, no. 6, pp. 1427–1439, 2006.

[23] T. Kuhn, M. Perc, and D. Helbing, "Inheritance patterns in citation networks reveal scientific memes," *Physical Review X*, vol. 4, no. 4, p. 041036, 2014.

[24] D. Zhao and A. Strotmann, "The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis," *Journal of the Association for Information Science and Technology*, vol. 65, no. 5, pp. 995–1006, 2014.

[25] I. G. Councill, C. L. Giles, and M.-Y. Kan, "Parscit: an open-source crf reference string parsing package." in *LREC*, 2008.

[26] C. Clark and S. Divvala, "Pdffigures 2.0: Mining figures from research papers," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, 2016, pp. 143–152.

[27] S. Tuarob, S. Bhatia, P. Mitra, and C. L. Giles, "Algorithmseer: A system for extracting and searching for algorithms in scholarly big data," *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 3–17, 2016.

[28] S. R. Choudhury, P. Mitra, A. Kirk, S. Szep, D. Pellegrino, S. Jones, and C. L. Giles, "Figure metadata extraction from digital documents," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 135–139.

[29] S. R. Choudhury, S. Tuarob, P. Mitra, L. Rokach, A. Kirk, S. Szep, D. Pellegrino, S. Jones, and C. L. Giles, "A figure search engine architecture for a chemistry digital library," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2013, pp. 369–370.

[30] S. Lawrence and C. L. Giles, "Searching the web: General and scientific information access," in *Internet Technologies and Services, 1999. Proceedings. First IEEE/Popov Workshop on*. IEEE, 1999, pp. 18–31.

[31] Á. Tejeda-Lorente, C. Porcel, E. Peis, R. Sanz, and E. Herrera-Viedma, "A quality based recommender system to disseminate information in a university digital library," *Information Sciences*, vol. 261, pp. 52–69, 2014.

[32] N. Barbuti, S. Ferilli, D. Redavid, and T. Caldarola, "An integrated management system for multimedia digital library," *Procedia Computer Science*, vol. 38, pp. 128–132, 2014.

[33] J. Sun and B.-Z. Yuan, "Development and characteristic of digital library as a library branch," *IERI Procedia*, vol. 2, pp. 12–17, 2012.

[34] J. Wu, K. Williams, H.-H. Chen, M. Khabsa, C. Caragea, A. Ororbia, D. Jordan, and C. L. Giles, "Citeseerx: Ai in a digital library search engine." in *AAAI*, 2014, pp. 2930–2937.

[35] A. E. Lovasz, E.-C. Lovasz, and C. M. Gruescu, "Digital library of mechanisms," *Procedia-Social and Behavioral Sciences*, vol. 163, pp. 85–91, 2014.

[36] F. Huang, J. Li, J. Lu, T. W. Ling, and Z. Dong, "Pandasearch: a fine-grained academic search engine for research documents."

[37] D. Lee, H.-s. Kim, E. K. Kim, S. Yan, J. Chen, and J. Lee, "Leedeo: Web-crawled academic video search engine," in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*. IEEE, 2008, pp. 497–502.

[38] J. DiMicco, D. R. Millen, W. Geyer, C. Dugan, B. Brownholtz, and M. Muller, "Motivations for social networking at work," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 2008, pp. 711–720.

[39] R. Van Noorden, "Online collaboration: Scientists and the social network." *Nature*, vol. 512, no. 7513, pp. 126–129, 2014.

[40] M. Thelwall and K. Kousha, "Academia. edu: social network or academic network?" *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 721–731, 2014.

[41] O. Almousa, "Users' classification and usage-pattern identification in academic social networks," in *Applied Electrical Engineering and Computing Technologies (AEECT), 2011 IEEE Jordan Conference on*. IEEE, 2011, pp. 1–6.

[42] M. Thelwall and K. Kousha, "Researchgate: Disseminating, communicating, and measuring scholarship?" *Journal of the Association for Information Science and Technology*, vol. 66, no. 5, pp. 876–889, 2015.

[43] V. Gewin, "Networking in vivo: An interdisciplinary networking site for scientists," *Nature*, vol. 462, no. 123, p. 4, 2009.

[44] D. MacMillan, "Data sharing and discovery: What librarians need to know," *The Journal of Academic Librarianship*, vol. 40, no. 5, pp. 541–549, 2014.

[45] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011.

[46] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[47] N. Khan, I. Yaqoob, I. A. T. Hashem, Z. Inayat, W. K. Mahmoud Ali, M. Alam, M. Shiraz, and A. Gani, "Big data: survey, technologies, opportunities, and challenges," *The Scientific World Journal*, vol. 2014, 2014.

[48] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *Access, IEEE*, vol. 2, pp. 652–687, 2014.

[49] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, "Defining and identifying sleeping beauties in science," *Proceedings of the National Academy of Sciences*, p. 201424329, 2015.

[50] M. E. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5200–5205, 2004.

[51] N. Akhtar, "Social network analysis tools," in *Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on*. IEEE, 2014, pp. 388–392.

[52] C. Chen, F. Ibekwe-Sanjuan, and J. Hou, "The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis," *CoRR*, vol. abs/1002.1985, 2010. [Online]. Available: http://arxiv.org/abs/1002.1985

[53] M. Bastian, S. Heymann, M. Jacomy *et al.*, "Gephi: an open source software for exploring and manipulating networks." *ICWSM*, vol. 8, pp. 361–362, 2009.

[54] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[55] V. Batagelj and A. Mrvar, "Pajek-program for large network analysis," *Connections*, vol. 21, no. 2, pp. 47–57, 1998.

[56] D. A. Schult and P. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, vol. 2008, 2008, pp. 11–16.

[57] E. Yan and Y. Ding, "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 7, pp. 1313–1326, 2012.

[58] F. Osareh, R. Khademi, M. K. Rostami, and M. S. Shirazi, "Co-authorship network structure analysis of iranian researchers scientific outputs from 1991 to 2013 based on the social science citation index (ssci)," *Collnet Journal of Scientometrics and Information Management*, vol. 8, no. 2, pp. 263–271, 2014.

[59] H. B. Andrade, E. de Los Reyes Lopez, and T. B. Martín, "Dimensions of scientific collaboration and its contribution to the

academic research groups' scientific quality," *Research Evaluation*, vol. 18, no. 4, pp. 301–311, 2009.

[60] A. Sigogneau, O. Malagutti, M. Crance, and S. Bauin, "Cross-disciplinary research: co-evaluation and co-publication practices of the cnrs laboratories," *Research evaluation*, vol. 14, no. 2, pp. 165–176, 2005.

[61] W. Glänzel and A. Schubert, "Analysing scientific networks through co-authorship," in *Handbook of quantitative science and technology research*. Springer, 2005, pp. 257–276.

[62] W. Glanzel and A. Schubert, "Domesticity and internationality in co-authorship, references and citations," *Scientometrics*, vol. 65, no. 3, pp. 323–342, 2005.

[63] J. Hoekman, K. Frenken, and R. J. Tijssen, "Research collaboration at a distance: Changing spatial patterns of scientific collaboration within europe," *Research Policy*, vol. 39, no. 5, pp. 662–673, 2010.

[64] R. Veugelers and B. Cassiman, "R&d cooperation between firms and universities. some empirical evidence from belgian manufacturing," *International Journal of Industrial Organization*, vol. 23, no. 5, pp. 355–379, 2005.

[65] A. Lowrie and P. J. McKnight, "Academic research networks:: A key to enhancing scholarly standing," *European Management Journal*, vol. 22, no. 4, pp. 345–360, 2004.

[66] Y. Ding, "Applying weighted pagerank to author citation networks," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 2, pp. 236–245, 2011.

[67] E. Yan and Y. Ding, "Weighted citation: An indicator of an article's prestige," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 8, pp. 1635–1643, 2010.

[68] L. Leydesdorff, "How are new citation-based journal indicators adding to the bibliometric toolbox?" *Journal of the American Society for Information Science and Technology*, vol. 60, no. 7, pp. 1327–1336, 2009.

[69] S. X. Zhao and F. Y. Ye, "Power-law link strength distribution in paper cocitation networks," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 7, pp. 1480–1489, 2013.

[70] C. Chen, F. Ibekwe-SanJuan, and J. Hou, "The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 7, pp. 1386–1409, 2010.

[71] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?" *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2389–2404, 2010.

[72] S. Milojević, C. R. Sugimoto, E. Yan, and Y. Ding, "The cognitive structure of library and information science: Analysis of article title words," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 10, pp. 1933–1953, 2011.

[73] X. Wang, J. Wang, F. Ma, and C. Hu, "The" small-world" characteristic of author co-words network," in *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*. IEEE, 2007, pp. 3717–3720.

[74] E. Yan, Y. Ding, and C. R. Sugimoto, "P-rank: An indicator measuring prestige in heterogeneous scholarly networks," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 3, pp. 467–477, 2011.

[75] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*. IEEE, 2011, pp. 121–128.

[76] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (kdt)." in *KDD*, vol. 95, 1995, pp. 112–117.

[77] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*. New York, NY, USA: Cambridge University Press, 2014.

[78] C. C. Aggarwal and C. Zhai, "An introduction to text mining," in *Mining text data*. Springer, 2012, pp. 1–10.

[79] M. Song and T. Chambers, "Text mining with the stanford corenlp," in *Measuring Scholarly Impact*. Springer, 2014, pp. 215–234.

[80] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining." in *Ldv Forum*, vol. 20, no. 1, 2005, pp. 19–62.

[81] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[82] J. Lin and D. Demner-Fushman, "Semantic clustering of answers to clinical questions," in *AMIA Annual Symposium Proceedings*, vol. 2007. American Medical Informatics Association, 2007, p. 458.

[83] W. Ke, C. R. Sugimoto, and J. Mostafa, "Dynamicity vs. effectiveness: studying online clustering for scatter/gather," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 19–26.

[84] F. Janssens, J. Leta, W. Glänzel, and B. De Moor, "Towards mapping library and information science," *Information processing & management*, vol. 42, no. 6, pp. 1614–1642, 2006.

[85] L. Leydesdorff and I. Hellsten, "Metaphors and diaphors in science communication mapping the case of stem cell research," *Science communication*, vol. 27, no. 1, pp. 64–99, 2005.

[86] R. N. Kostoff, J. A. del Rio, J. A. Humenik, E. O. Garcia, and A. M. Ramirez, "Citation mining: Integrating text mining and bibliometrics for research user profiling," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 13, pp. 1148–1156, 2001.

[87] A. Porter, A. Kongthon, and J.-C. Lu, "Research profiling: Improving the literature review," *Scientometrics*, vol. 53, no. 3, pp. 351–370, 2002.

[88] H. Kim and J. Lee, "Archiving research trends in lis domain using profiling analysis," *Scientometrics*, vol. 80, no. 1, pp. 75–90, 2009.

[89] S. Bhattacharya, H. Kretschmer, and M. Meyer, "Characterizing intellectual spaces between science and technology," *Scientometrics*, vol. 58, no. 2, pp. 369–390, 2003.

[90] R. Li, T. Chambers, Y. Ding, G. Zhang, and L. Meng, "Patent citation analysis: Calculating science linkage based on citing motivation," *Journal of the Association for Information Science and Technology*, vol. 65, no. 5, pp. 1007–1017, 2014.

[91] P. Glenisson, W. Glänzel, and O. Persson, "Combining full-text analysis and bibliometric indicators. a pilot study," *Scientometrics*, vol. 63, no. 1, pp. 163–180, 2005.

[92] M. Song and S. Y. Kim, "Detecting the knowledge structure of bioinformatics by mining full-text collections," *Scientometrics*, vol. 96, no. 1, pp. 183–201, 2013.

[93] X. Liu, J. Zhang, and C. Guo, "Full-text citation analysis: enhancing bibliometric and scientific publication ranking," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 1975–1979.

[94] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[95] B. Hu, X. Dong, C. Zhang, T. D. Bowman, Y. Ding, S. Milojević, C. Ni, E. Yan, and V. Larivière, "A lead-lag analysis of the topic evolution patterns for preprints and publications," *Journal of the Association for Information Science and Technology*, 2015.

[96] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 113–120.

[97] J. Tang, R. Jin, and J. Zhang, "A topic modeling approach and its integration into the random walk framework for academic search," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 1055–1060.

[98] Y. Ding, "Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks," *Journal of informetrics*, vol. 5, no. 1, pp. 187–203, 2011.

[99] Y. Dong, R. A. Johnson, and N. V. Chawla, "Can scientific impact be predicted?" *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 18–30, 2016.

[100] J. Zhang, F. Xia, W. Wang, X. Bai, S. Yu, T. M. Bekele, and Z. Peng, "Cocarank: A collaboration caliber-based method for finding academic rising stars," in *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 395–400.

[101] J. Zhang, Z. Ning, X. Bai, W. Wang, S. Yu, and F. Xia, "Who are the rising stars in academia?" in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, 2016, pp. 211–212.

[102] C.-T. Zhang, "A novel triangle mapping technique to study the h-index based citation distribution," *Scientific reports*, vol. 3, 2013.

[103] D. Walker, H. Xie, K.-K. Yan, and S. Maslov, "Ranking scientific publications using a model of network traffic," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, no. 06, p. P06010, 2007.

[104] H. Sayyadi and L. Getoor, "Futurerank: Ranking scientific articles by predicting their future pagerank." in *SDM*. SIAM, 2009, pp. 533–544.

[105] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[106] L. Egghe, "Theory and practise of the g-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.

[107] D. Zhou, S. Orshanskiy, H. Zha, C. L. Giles *et al.*, "Co-ranking authors and documents in a heterogeneous network," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 739–744.

[108] Q. Meng and P. J. Kennedy, "Discovering influential authors in heterogeneous academic networks by a co-ranking method," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 1029–1036.

[109] Z. Liu, H. Huang, X. Wei, and X. Mao, "Tri-rank: An authority ranking framework in heterogeneous academic networks by mutual reinforce," in *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*. IEEE, 2014, pp. 493–500.

[110] A. Fersht, "The most influential journals: Impact factor and eigenfactor," *Proceedings of the National Academy of Sciences*, vol. 106, no. 17, pp. 6883–6884, 2009.

[111] C. T. Bergstrom, J. D. West, and M. A. Wiseman, "The eigenfactor? metrics," *The Journal of Neuroscience*, vol. 28, no. 45, pp. 11 433–11 434, 2008.

[112] F. Xia, H. Liu, I. Lee, and L. Cao, "Scientific article recommendation: Exploiting common author relations and historical preferences," *IEEE Transactions on Big Data*, vol. PP, no. 99, pp. 1–1, 2016.

[113] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 421–430.

[114] M. Gori and A. Pucci, "Research paper recommender systems: A random-walk based approach," in *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*. IEEE, 2006, pp. 778–781.

[115] O. Kucuktunc, K. Kaya, E. Saule, and U. V. Catalyurek, "Fast recommendation on bibliographic networks," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 2012, pp. 480–487.

[116] N. Y. Asabere, F. Xia, Q. Meng, F. Li, and H. Liu, "Scholarly paper recommendation based on social awareness and folksonomy," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 30, no. 3, pp. 211–232, 2015.

[117] F. Xia, N. Y. Asabere, H. Liu, N. Deonauth, and F. Li, "Folksonomy based socially-aware recommendation of scholarly papers for conference participants," in *Proceedings of the companion publication of the 23rd international conference on World wide web companion*. International World Wide Web Conferences Steering Committee, 2014, pp. 781–786.

[118] S. Lee and B. Bozeman, "The impact of research collaboration on scientific productivity," *Social studies of science*, vol. 35, no. 5, pp. 673–702, 2005.

[119] X. Su, W. Wang, S. Yu, C. Zhang, T. M. Bekele, and F. Xia, "Can academic conferences promote research collaboration?" in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, 2016, pp. 231–232.

[120] X. Kong, H. Jiang, Z. Yang, Z. Xu, F. Xia, and A. Tolba, "Exploiting publication contents and collaboration networks for collaborator recommendation," *PloS one*, vol. 11, no. 2, p. e0148492, 2016.

[121] G. R. Lopes, M. M. Moro, L. K. Wives, and J. P. M. De Oliveira, "Collaboration recommendation on academic social networks," in *Advances in Conceptual Modeling–Applications and Challenges*. Springer, 2010, pp. 190–199.

[122] M. A. Brandão, M. M. Moro, G. R. Lopes, and J. P. Oliveira, "Using link semantics to recommend collaborations in academic social networks," in *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013, pp. 833–840.

[123] F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, "Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors," *Emerging Topics in Computing, IEEE Transactions on*, vol. 2, no. 3, pp. 364–375, 2014.

[124] J. Li, F. Xia, W. Wang, Z. Chen, N. Y. Asabere, and H. Jiang, "Acrec: a co-authorship based random walk model for academic collaboration recommendation," in *Proceedings of the companion publication of the 23rd international conference on World wide web companion*. International World Wide Web Conferences Steering Committee, 2014, pp. 1209–1214.

[125] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1285–1293.

[126] Y. Guo and X. Chen, "Cross-domain scientific collaborations prediction using citation," in *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*. IEEE, 2013, pp. 765–770.

[127] M. C. Pham, D. Kovachev, Y. Cao, G. M. Mbogos, and R. Klamma, "Enhancing academic event participation with context-aware and social recommendations," in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. IEEE, 2012, pp. 464–471.

[128] M. F. Hornick and P. Tamayo, "Extending recommender systems for disjoint user/item sets: The conference recommendation problem," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 8, pp. 1478–1490, 2012.

[129] F. Xia, N. Y. Asabere, H. Liu, Z. Chen, and W. Wang, "Socially aware conference participant recommendation with personality traits," 2014.

[130] N. Yaw Asabere, F. Xia, W. Wang, J. J. Rodrigues, F. Basso, and J. Ma, "Improving smart conference participation through socially aware recommendation," *Human-Machine Systems, IEEE Transactions on*, vol. 44, no. 5, pp. 689–700, 2014.

[131] P. R. Carlile, "Working knowledge: how organizations manage what they know," *People and Strategy*, vol. 21, no. 4, p. 58, 1998.

[132] N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins, "P@ noptic expert: Searching for experts not just for documents," in *Ausweb Poster Proceedings, Queensland, Australia*, vol. 15, 2001, p. 17.

[133] D. Yimam-Seid and A. Kobsa, "Expert-finding systems for organizations: Problem and domain analysis and the demoir approach," *Journal of Organizational Computing and Electronic Commerce*, vol. 13, no. 1, pp. 1–24, 2003.

[134] N. Craswell, A. P. de Vries, and I. Soboroff, "Overview of the trec 2005 enterprise track." in *Trec*, vol. 5, 2005, pp. 199–205.

[135] Y. Xu, X. Guo, J. Hao, J. Ma, R. Y. Lau, and W. Xu, "Combining social network and semantic concept analysis for personalized academic researcher recommendation," *Decision Support Systems*, vol. 54, no. 1, pp. 564–573, 2012.

[136] D.-R. Liu, Y.-H. Chen, W.-C. Kao, and H.-W. Wang, "Integrating expert profile, reputation and link analysis for expert finding in question-answering websites," *Information Processing & Management*, vol. 49, no. 1, pp. 312–329, 2013.

[137] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang, "Expertrank: A topic-aware expert finding algorithm for online knowledge communities," *Decision Support Systems*, vol. 54, no. 3, pp. 1442–1451, 2013.

[138] J. D. West, I. Wesley-Smith, and C. T. Bergstrom, "A recommendation system based on hierarchical clustering of an article-level citation network," 2016.

[139] C.-J. Wu, J.-M. Chung, C.-Y. Lu, H.-M. Lee, and J.-M. Ho, "Using web-mining for academic measurement and scholar recommendation in expert finding system," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2011, pp. 288–291.

[140] M. Neshati, S. H. Hashemi, and H. Beigy, "Expertise finding in bibliographic network: Topic dominance learning approach," *Cybernetics, IEEE Transactions on*, vol. 44, no. 12, pp. 2646–2657, 2014.

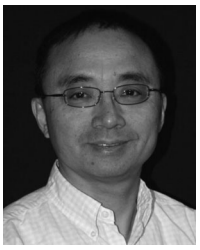
[141] X. Bai, F. Xia, I. Lee, J. Zhang, and Z. Ning, "Identifying anomalous citations for objective evaluation of scholarly article impact," *PloS one*, vol. 11, no. 9, p. e0162364, 2016.

[142] H. Piwowar, "Altmetrics: Value all research products," *Nature*, vol. 493, no. 7431, pp. 159–159, 2013.

FENG XIA (M07-SM12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He was a Research Fellow with the Queensland University of Technology, Australia. He is currently a Full Professor with the School of Software, Dalian University of Technology, China. He is the (Guest) Editor of several international journals. He serves as the General Chair, PC Chair, Workshop Chair, or Publicity Chair of a number of conferences. He has authored two books and over 200 scientific papers in international journals and conferences. His research interests include social computing, computational social science, big data, and mobile social networks. He is a Senior Member of IEEE and ACM.

WEI WANG received his B.S. degree in Electronic Information Science and Technology from Shenyang University, Shenyang, China, in 2012. He is currently working toward the Ph.D. degree in Software Engineering at Dalian University of Technology, Dalian, China. His research interests include big scholarly data, social network analysis and computational social science.

TESHOME MEGERSA BEKELE received the B.Sc. degree in computer science from Hawassa University, Awassa, Ethiopia, in 2006 and the M.Eng. degree in software engineering from Chongqing University, Chongqing, China, in 2011. He is currently pursuing the Ph.D. degree with the School of Software, Dalian University of Technology, Dalian, China. His current research interests include computational social science, big scholarly data and recommender systems.



HUAN LIU received the BEng degree in computer science and electrical engineering from Shanghai JiaoTong University, Shanghai, China, and the PhD degree in computer science from the University of Southern California, Los Angeles, CA. He is currently a professor of computer Science and Engineering at Arizona State University (ASU), Tempe, AZ. Before he joined ASU, he worked at Telecom Australia Research Labs and was on the faculty at the National University of Singapore. He was recognized for excellence in teaching and research in Computer Science and Engineering at Arizona State University. His research interests are in data mining, machine learning, social computing, and artificial intelligence, investigating problems that arise in many real-world, data-intensive applications with high-dimensional data of disparate forms such as social media. His well-cited publications include books, book chapters, encyclopedia entries as well as conference and journal papers. He serves on journal editorial boards and numerous conference program committees, and is a founding organizer of the International Conference Series on Social Computing, Behavioral-Cultural Modeling, and Prediction (http://sbp.asu.edu/). He is a Fellow of the IEEE and an ACM Distinguished Scientist.