IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Survey of Scholarly Data Visualization

JIAYING LIU[1], TAO TANG[2],WEI WANG[1], BO XU[1], XIANGJIE KONG[1], (SENIOR MEMBER, IEEE), AND FENG XIA[1], (SENIOR MEMBER, IEEE)

[1]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China.
[2]Chengdu College, University of Electronic Science and Technology of China, Chengdu 611731, China.

Corresponding author: F. Xia (e-mail:f.xia@ieee.org).

ABSTRACT Scholarly information usually contains millions of raw data such as authors, papers, citations, as well as scholarly networks. With the rapid growth of the digital publishing and harvesting, how to visually present the data efficiently becomes challenging. Nowadays, various visualization techniques can be easily applied on scholarly data visualization and visual analysis, which enables scientists to have a better way to represent the structure of scholarly datasets and reveal hidden patterns in the data. In this paper, we first introduce the basic concepts and the collection of scholarly data. Then we provide a comprehensive overview of related data visualization tools, existing techniques as well as systems for analyzing volumes of diverse scholarly data. Finally, open issues are discussed to pursue new solutions for abundant and complicated scholarly data visualization as well as techniques that support a multitude of facets.

INDEX TERMS Scholarly Data, scholarly data analysis, scholarly data visualization, visual analysis.

I. INTRODUCTION

SCHOLARLY data contain abundant academic resources such as scholarly documents (i.e., papers, books, patents, and scientific reports) as well as associated data including information of authors, citations, figures, tables, etc. In addition, the concept of scholarly networks is central to the study of scholarly data [1]. Scholarly data have become a vital part of scientific research with the appearance of various digital libraries and the rapid development of scholarly data analysis technologies [2]. This enables researchers to look into the science itself with a new angle by studying scholarly data [3], [4]. Beyond that, it also helps researchers to learn better about knowledge production processes with wide varieties of scholarly data analysis methods. Meanwhile, the analysis of scholarly data is not only important for academia, but also promotes the understanding of human social activities, i.e., for sociologists to observe researcher interactions [5] and community formation [6], for countries to evaluate the impact of institutions or scientists and to allocate resources [7].

Along with driving key insights through analyzing scholarly data, visually presenting the data is also crucial. Visualization can be described as “make something visible". To be more specific, it represents the process or ability to form a sensible mental picture in a person’s brain. It can also serve as a target to express the visualization of results [8]. In the era of the proliferation of information technology, visualization means the technology which can enhance human’s perception by using perceptual competence to visualize the interaction of data [9]. It not only converts the raw data into intuitional graphics, symbols, colors, arts, etc., but also enhances data recognition efficiency and passes valid information. In a word, data visualization transforms the data into multiple easy-understand forms. Initially, two well-established branches of data visualization are that of scientific data visualization and unstructured information visualization. As data analysis is becoming more and more important [10], combining visualization with analytics, which is called visual analytics, becomes one of the major areas of interest within the field of data visualization. Based on the theory of computer graphics, scientific visualization aims to create a visual expression instead of numerical complex scientific concepts or results [11]. In comparison with scientific visualization, information visualization focuses on dealing with the unstructured high-dimensional data such as textual data, financial data, and multimedia data [12].

The process of scholarly data visualization combines the theory of scientific visualization, information visualization, and visual analytics. It can transform the scholarly dataset into an appropriate representation. It also enables scientists to have a better understanding of the structure and dynamics of science through a visual way. Visualization technologies have a strong applicability in scholarly data. It is suitable for users to visualize the data in different ways by using tools with or without a programming language. For instance, based

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/ACCESS.2018.2815030, IEEE Access

on the minimum spanning tree, visualization of the author can reflect the collaboration network as a two-dimensional picture. At the same time, it provides a key insight into the direct influence on authors [13].

Scholarly data analysis tries to deal with problems within the scope of Science of Science [14], [15]. However, as more scholarly data are available to scientists, how to use the sheer quantity data becomes a critical problem to be solved [16]. Fortunately, with the development of visualization technologies, it is much easier to have a better understanding of scholarly data [17] and put it to great use [18], [19]. For example, it can help researchers to understand how scientists interact with each other. It also enables researchers to mine implicit relationships which are hidden in citation networks especially in co-citation networks [20]. Among data visualization tools, CiteSpace [21] is well known by its strong ability to analyze the co-citation network. Through the visualization of Institute for Scientific Information (ISI), Chinese Social Sciences Citation Index (CSSCI), China National Knowledge Infrastructure (CNKI), and the analysis of the other literature databases, CiteSpace can help to keep track of research areas' hot spots and trends. It helps researchers to understand the research frontier and the evolution of critical pathways, important literature, authors, and institutions as well.

The purpose of this paper is to review recent research into the field of scholarly data visualization. To the best of our acknowledge, although scholarly data visualization is important, there is no study that provides a comprehensive review of it. There are two primary aims of this study: 1) To provide a comprehensive understanding of development and challenges in the field of scholarly data visualization. 2) To find significant issues which can ascertain the future of this emerging discipline. Therefore, we summarize the overall research issues on scholarly data visualization. The main issues addressed in this paper include: details of scholarly data and visualization technologies, visualization of scholarly data including the single attribute and heterogeneous networks in academia, and visual analytic systems of scholarly data. The scholarly data visualization section is concerned with the tools and systems used for scholarly data presentation and analysis.

The remaining part of the paper proceeds as follows. Details of scholarly data and collection methods are presented in Section II. Section III lays out theoretical dimensions of the visualization tools. Section IV is concerned with the generic visualization tools and systems used for scholarly data. Section V shows how visualization can be combined with various analytical techniques in different visual analytic systems to enlarge the understanding of scholarly data. Finally, this survey is concluded and the future work is highlighted in Section VI. The overall idea of exploring scholarly data visualization is summarized in FIGURE 1.

II. SCHOLARLY DATA COLLECTION

Before the visualization of scholarly big data, it is crucial to collect and manage the relevant scholarly datasets. In the age of big data and open science, more and more scholarly documents can be freely accessed from the Internet. In this section, we introduce the entities in scholarly datasets and how to collect them for data visualization.

A. SCHOLARLY DATA EXTRACTION

Scholarly data are clearly heterogeneous with various entities. Sinha et al. [22] model the academic community as a heterogeneous graph consisting of six types of entities including author, paper, venue, institution, event, and field of study. Specifically, the venue entity means the journal and conference series, e.g., WWW, KDD, TKDE, etc. The event entity means the conference instances, e.g., KDD 2017. The goal of scholarly data visualization is to present the dynamic relationships vividly among these different entities. Most of these entities can be inferred from the raw data, e.g., author, and citation information. We describe how to collect these data entities from scholarly dataset, e.g., DBLP in the following subsections.

1) Raw Data Extraction

Raw data (or Metadata) is the first set of data extracted from the online digital libraries or academic search engines. It is the basis for academic searching, indexing, and visualization. Specifically, raw data contains authors, title, abstract, keywords, venue, publisher, page number, date of publication, DOI, etc. In order to collect raw data, rule-based metadata extraction has been proposed. For example, Guo and Jin [23] propose a rule-based framework for metadata extraction from scientific papers. The framework utilizes format information such as font size and position to guide the metadata extraction process. They use header information as rules which contain author, title, abstracts located in the first page of the paper. Such rule-based raw data extraction techniques can achieve high accuracy in header information extraction.

2) Author Information Extraction

Most scholarly data visualization requires author information for analysis [24]. Usually, an article can be used to trace user information. The author entity can be well profiled with article raw data including the author's name, affiliation, research interest, etc. Furthermore, based on the author information, the most important academic relationship, co-authorship, can be inferred, where two authors are considered to be connected if they have co-authored a common paper. Author entity is usually the basis of academic search engines and digital libraries. CiteseerX [25] proposes to infer author's information including name, institution, affiliation, and email address from the PDF-converted documents. However, many aspects of author information are neglected if we merely focus on the article itself. Yao et al. [26] propose to build a semantic profile for an academic researcher by identifying and annotating author information on the Web including academic search engines and authors' homepages. Their approach has been proven to achieve a significant improvement

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/ACCESS.2018.2815030, IEEE Access

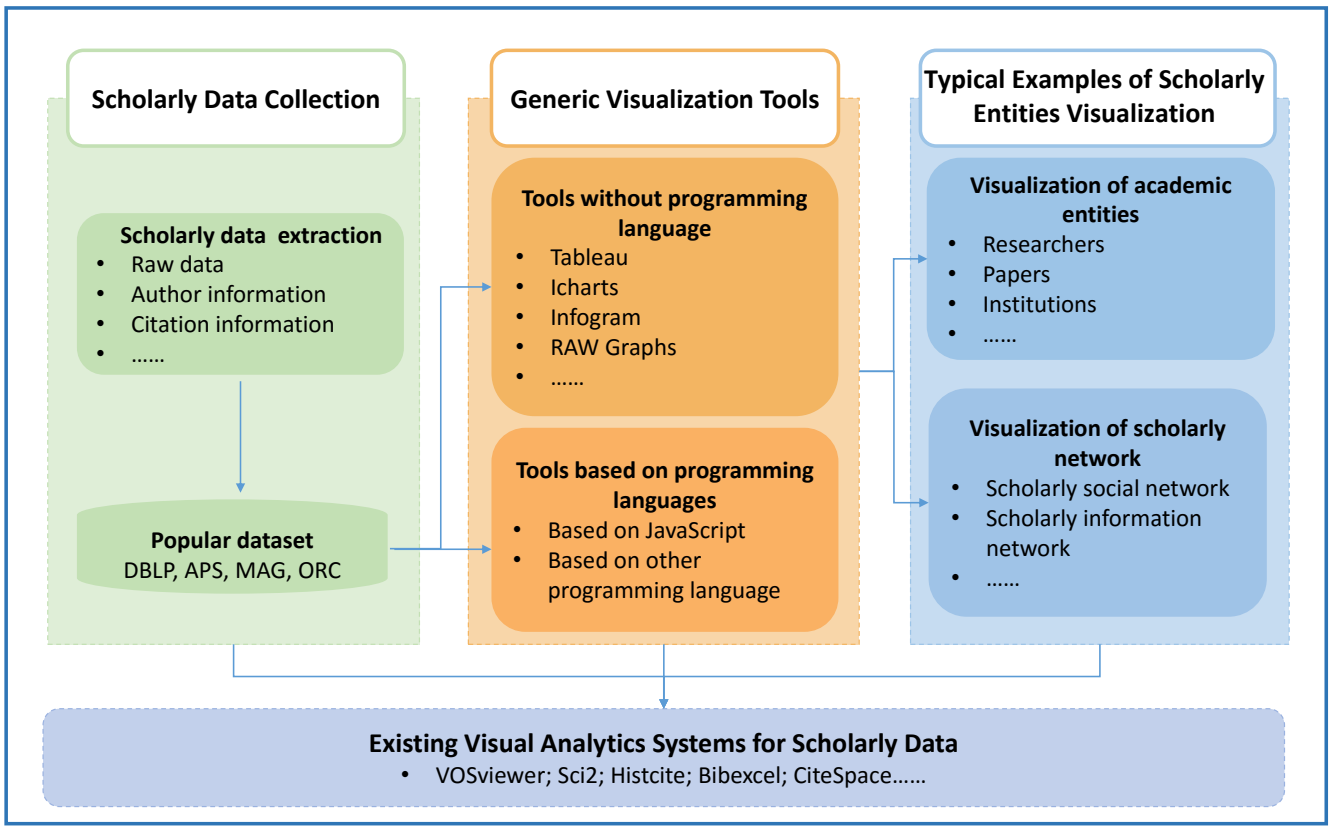

FIGURE 1: Framework of scholarly data visualization.

in identifying expert. Meanwhile, a scholar-centric academic search engine Aminer [27] is built based on this approach.

3) Citation Information Extraction

Another important relationship in scholarly big data is the citation relationship. If paper A appears in the reference list of paper B, it means that paper B cites the paper A. The citation relationship is a direct relationship which has been extensively used to quantify academic impact as well as to trace the origin of new knowledge. Since the citation information is located in the “Reference" section of a paper, citation information extraction requires accurately locating a section of a given paper with indicator “References", “Bibliography", or “Sources". Usually, these section can be found at the end of a given paper. ParsCit [28], FLUX-CiM [29], and CRF-based system [30] are the three widely-used tools to extract citation information. Bai et al. [31] extract citations of Physical Review C (PRC) from the whole APS dataset, then they extract the co-author COI relationship and the suspected COI relationship for obtaining the existing citation relationship between every two articles. Furthermore, they propose an efficient method for identifying anomalous citations for objective evaluation of the scholarly article impact.

4) Other Information Extraction

There are various other entities in the scholarly big data including the institutions, venues, as well as the content information like algorithms and figures. The Microsoft Academic Search [22] proposes to collect venue entities from a few semi-structured websites from Bing which are used by conference organizers to post conference information. They further conflate the venue events including conference instances and series across different websites with various signals inferred from the semi-structured data including full name, year, location, etc. Meanwhile, the journal and institution entities are mainly extracted from the in-house knowledge bases, for example, ACM digital library.

Apart from these mentioned entities, there are various other entities and knowledge which need to be further investigated. For example, figures, tables, algorithms, and the acknowledgement are also important to analyze and visualize scholarly big data. Figures are usually able to vividly present the architecture of a paper and most results are presented with figures or tables. Algorithms are used to describe the core process and idea of the proposed method. The Acknowledgement section usually indicates scholars who also contribute to the paper and grants information. The extraction of these scholarly data types can support the scholarly data visualization with additional information. However, existing works in collecting these data are limited to raw data extraction [32].

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/ACCESS.2018.2815030, IEEE Access

TABLE 1: Statistics of popular scholarly dataset.

Name	Size	Field	Citation
DBLP	382 MB	Computer Science	No
APS	1.2 GB	Physics	Yes
MAG	29.8 GB	Multidisciplinary	Yes
ORC	7 GB	Computer Science/Neuroscience	Yes

B. POPULAR DATASET

In modern academia, more and more scholars are willing to share their datasets. Many academic search engines, digital libraries, and research institutions have made their scholarly datasets available to help explore science itself. Among them, DBLP, APS (American Physical Society), MAG (Microsoft Academic Graph), and ORC (Open Research Corpus) are widely used. These datasets are extracted from the publication information of a certain discipline or various disciplines contain the basic information of a given scholar. The statistics of DBLP dataset can be seen from FIGURE 2. The basic information of these datasets can be seen in TABLE 1.

FIGURE 2: Statistics of DBLP digital library.

1) DBLP

DBLP digital library is an online reference bibliographic information focusing on the field of Computer Science area. It contains over 3.8 million publications that are published by more than 1.9 million authors. It indexes more than 32,000 journal volumes and more than 31,000 conferences or workshop proceedings. More importantly, each author name is uniquely indexed, making DBLP a fully-fledged freely available properly disambiguate large datasets to date. They have given an additional number for scholars with the same name, e.g, "Wei Wang 0077". However, one of the limitations of the dataset is that it does not contain the citation information. The dataset can be downloaded directly from the link http://dblp.dagstuhl.de/xml/ in the formation of .xml data.

2) APS

APS dataset is provided by the American Physical Society, which is a corpus of Physical Review Letter, Physical Review, and reviews of Modern Physics. It is composed of more than 450,000 articles dating back to 1893. Scholars may now request access to the APS dataset from the link https://journals.aps.org/datasets with accepting the terms and conditions governing the use of the datasets. This dataset contains two sub-datasets including article metadata and citing article pairs. The article metadata contains the raw data of all APS journal articles including the DOI, the venue, the page number or the first and the last page, the title, authors, affiliations, etc. The citing article dataset contains pairs of APS articles that cite each other. The dataset is in the format of comma-separated values (CSV).

3) MAG

The MAG dataset is a heterogeneous academic graph containing scientific publication records and the citation relationships. It mainly consists of six entity types including authors, papers, institutions, journals, conferences, and the field of study. The dataset now can be accessed via the Microsoft Cognitive Services Academic Knowledge API with the link http://research.microsoft.com/en-us/projects/mag/. The biggest advantage of this dataset is that it contains the papers from all fields. However, name disambiguation is necessary before utilizing this dataset for further analysis.

4) ORC

The ORC dataset is provided by the Semantic Scholar project. It contains more than 7 million papers from the fields of Computer Science and Neuroscience. It contains the raw data including the paper title, abstract, keywords, the paper's url, authors' name, in and out citation, the publication date, and the venue.

III. GENERIC VISUALIZATION TOOLS

With the development of visualization tools, most of them have integrated many useful functions (e.g., data preprocessing, visual analysis) into a library or common plugins. Therefore, it enables users to simplify the procedure of data visualization by using a programming language to invoke the function or using the pre-integrated function directly in tools. Visualization tools also give users ability to transform every element of the data into interactive charts and pictures [33]. Based on these intuitive charts and pictures, the analysis of the generated charts and graphs is more effective than the raw data. Data visualization is frequently used in BI (Business Intelligence), scientific visualization, information visualization, and visual analysis. Therefore, when the researchers handle the large, complex datasets, it demands the traditional data mining techniques with high levels of data processing. According to users' preference, we divide tools into two categories, visualization tools without the programming language and tools with a programming language.

A. TOOLS WITHOUT THE PROGRAMMING LANGUAGE

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/ACCESS.2018.2815030, IEEE Access

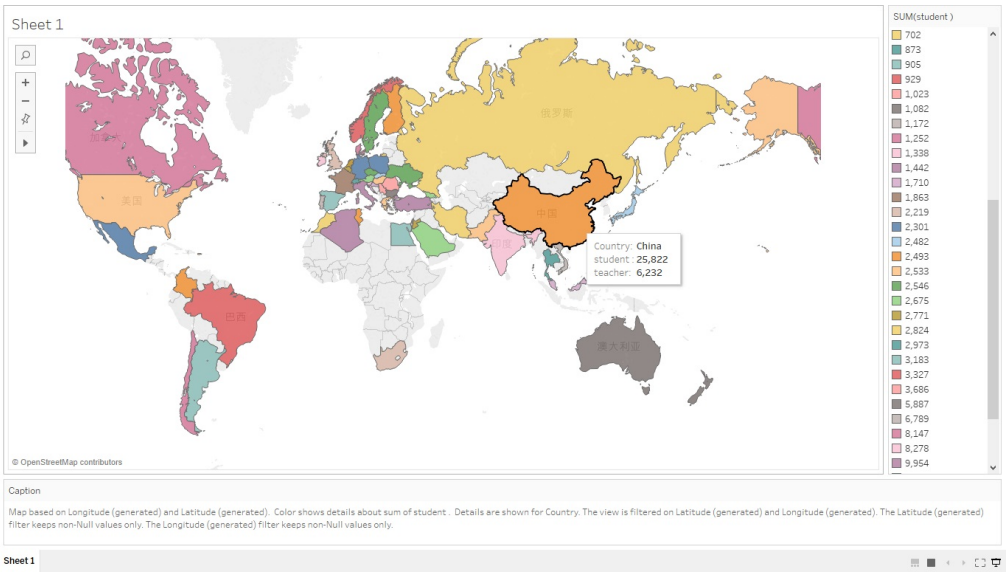

FIGURE 3: The number of graduate students and their supervisors between 63 countries on the world map which is generated by Tableau. It maps the students from different countries in different colors.

1) Tableau

Tableau is a desktop software for business intelligence and analytics. Tableau can be linked to the data files on both local and server. It supports a variety of data file formats such as txt, xlsx, and csv. It also has a number of database interfaces for importing data from online servers, such as Oracle and MySQL. The updated data in the server can synchronize with the local automatically. Tableau uses a novelty operation method. It extracts the header of each item in the linked or imported dataset automatically. The graph is immediately generated when users drag and drop these headers into row and column and choose a chart type. The map function in Tableau is easy to use. For example, the geographic information in users' datasets can be automatically marked and displayed on a map by using the map function. It is suitable for users to analyze the data geographically. In FIGURE 3, we use the map function in Tableau to visualize the data into a world map. Kale and Balan [34] draw some charts to analyze the job vacancy in New York by using Tableau.

Tableau contains several features: 1) It is easy to visualize the data, users only need to choose a chart type and drag the header after importing the data. 2) These flexible, interactive graphs allow users to analyze the characteristics of their data from various perspectives. 3) The public edition is free but it is commercial for using desktop edition, server edition, web-based edition, and even mobile application.

2) ICharts

ICharts[1] is a commercial web-based application that integrates the official optimized API connector for NetSuite, Salesforce, Google Cloud Platform, and many other database platforms. It is mainly used for BI. ICharts can combine data from Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), and even on-premise data storage that can help users take the comprehensive analysis of the data. ICharts declares itself as a real-time business intelligence tool because the databases it linked are automatically updated. It also provides a number of chart types for users to visualize their data, each type of the chart can be fully customized.

Some features of ICharts includes: 1) It connects to the real-time database, avoiding the secondary update of the data. 2) ICharts can take visual analysis automatically and build report of users' datasets periodically. The report is easy for sharing, thus other users can take visualization analysis individually. 3) ICharts can combine multiple different types of datasets into the dashboard by creating fully customized interactive charts.

3) Infogram

Infogram[2] is a web-based application for making graphs and charts about information, and this tool has a quick response and it can complete the data visualization quickly. Registered users could upload their own data file (.xls, .csv, .xlsx) to the website, as well as importing data on GoogleDrive, Dropbox, OneDrive or JSON feed. The problem of Infogram is that the project is created by a public URL, thus the data have no privacy. If users want to protect the privacy of their data, becoming a paid member is the only way. Infogram has opened its global sources which include all public themes and charts created by other users to let users share their inspiration to each other. Infogram also enhances the function of sharing, so that users could embed their charts on a web page by using the codes which are automatically generated or

[1] http://icharts.net

[2] https://infogram.com/

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/ACCESS.2018.2815030, IEEE Access

shared by URL and Email. This application is easy for users to visualize their data. In FIGURE 4, we create a graph based on Infogram. It is a word cloud of the advisor-advisee dataset. The larger size of the country name represents the larger total number of the advisors/advisees.

Infogram contains several features: 1) It shows a friendly user experience, users can communicate with the technical staff online, which make their work easier. 2) The public chart type library shared by other users is a good place to share charts and get original inspirations from other productions. 3) It also provides a real-time data processing, and supports multi-terminal display. 4) The uploaded data in this tool's online database is public unless the user upgrades to a paid member in order to make sure its privacy protection.

FIGURE 4: The word cloud of the advisor-advisee data visualized by Infogram. The size of the countries' name shows the total number of advisors/advisees.

4) RAW Graphs

RAW Graphs[3] is an open web-based tool that can be used directly without registration. It supports the data format such as .tsv, .dsv, .csv, .json or .xls file, even the online data with a public API or from a public cloud platform. RAW Graphs processes the data that only using the web browser in the local. It doesn't upload it to the server which can ensure the data safe. This application offers users 21 kinds of chart models for their data visualization and also supports to create custom vector-based visualizations on top of the D3.js library by Bostock at local[4]. In FIGURE 5, we visualize the advisor-advisee dataset into a circular dendrogram. Users can choose a chart type and map the dimensions by dragging the visual variables of the selected layout into its attributes for convenient chart generating. They can also export the generated chart as a vector (SVG) or raster (PNG) image, or embed their graphs into web pages by using codes generated in RAW Graphs automatically.

RAW Graphs contains some features: 1) It is easy for users to visualize their data in charts. 2) The imported data is safe because it is only processed by the web browser but not on the online server. 3) RAW Graphs is open for users to create new charts by D3.js, but it doesn't insert in web application for the customized charts.

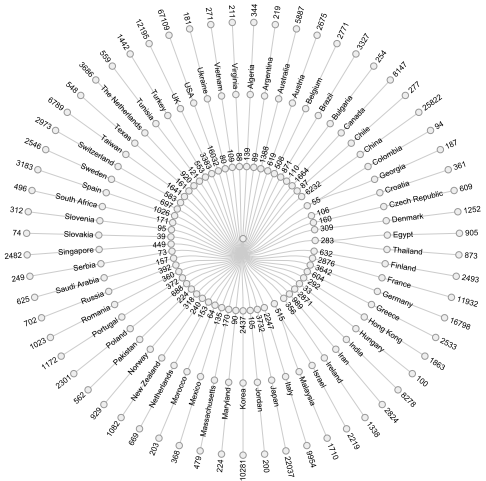

FIGURE 5: The number of advisors and advisees for different countries drawing by RAW Graphs. Circles from inside to outside represent: the number of advisors, countries, the number of advisees, respectively.

5) Visualize Free

Visualize Free[5] is a free and light web-based application which needs registration before using. Users can upload their data files with a 5MB file limit, and it supports Excel file (both .xls and .xlsx) and text file (.csv and tab-delimited.txt). Users can easily visualize their data with multiple beautiful charts by dragging and dropping the data into the correct layout for shaping the chart's dimension. The free visual analytics is provided that users can compile the detailed analysis of the uploaded data. In Visualize Free, the uploaded data is private for users, and the generated charts can be shared by moving them into the shared folder or downloading them in .pdf, .xls, or .ppt format.

Visualize Free contains some features: 1) It is easy for using and suitable for users to visualize the small quantity of data into exquisite charts, and visual analysis is free for analyzing their data. 2) Common charts and maps are available for visualizing the data and it is easy for users to take visual analysis based on them.

[3]https://rawgraphs.io/

[4]https://github.com/densitydesign/raw

[5]https://www.visualizefree.com/

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/ACCESS.2018.2815030, IEEE Access

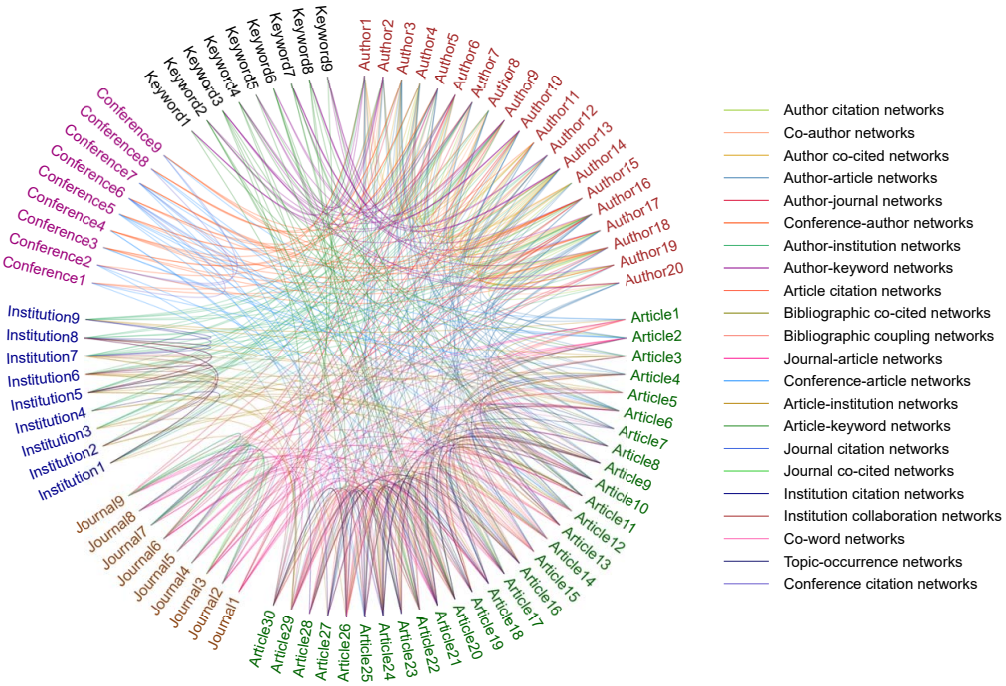

FIGURE 6: Different networks for various scholarly entities and their relationships visualized by using D3.js [35].

Based on visualization tools that we present above in this subsection, we visualize the same dataset into three different types of graphs (see in FIGURE 3, 4, 5). All tools we used support vector image format output, but this permission is only available for registered users in Tableau and subscriber in Infogram. Raw Graphs is totally free for users but it has less chart types.

B. TOOLS BASED ON A PROGRAMMING LANGUAGE

Visualization tools without the programming language are easy for users to visualize their data into common charts or graphs. Some visualization tools have opened their API which can enhance the function of chart plotting. It enables users to design their charts and graphs in a flexible way by handling the raw data with codes in own style, but it will cost time for beginners to command the new programming language. A part of the visualization tools are combined with JavaScript, other parts use programming languages such as Python, JAVA, PHP, and R. So we divide these visualization tools which are based on programming languages into two parts: tools based on JavaScript and tools based on other programming languages. Users can choose a tool and access to its official documents and relevant tutorials to learn how it works and practice to visualize the data into charts by themselves. In TABLE 2, We choose the five commonly used tools based on JavaScript for comparison, and the more detailed comparison of tools based on JavaScript is available in https://en.wikipedia.org/w/index.php?title=Comparison_of_JavaScript_charting_frameworks&oldid=805174557. In TABLE 3, we compare five visualization tools based on other programming languages.

1) Tools based on JavaScript

a: D3.js

D3.js [36] is a program of the open source JavaScript graphics library that combines HTML and CSS techniques, and the graphs it generated are all in .svg format after visualizing the imported data [37]. D3.js completes the data visualization that it runs as a coded html file in browser platform under a server environment. D3.js is requested to run with its official document library for function invocation. In the website[6], D3.js provides plenty of examples (i.e. graphs, charts and their source codes) for users which can inspire them to design their own charts or use the examples directly. FIGURE 6 is the display of different networks with various scholarly entities and their relationships by using D3.js [35]. In addition, D3.js can be well applied to geographic information display. For instance, FIGURE 7 is the coordinate of different institutions published papers on KDD conference in 2012 [38]. The red dots represent the institutions, and the green dot represents the location where the conference is held.

b: Chart.js

Chart.js[7] is also a program of the open source JavaScript graphics library. It uses canvas on HTML5, so the rendering performance is good in all modern browsers (above IE9). Chart.js can visualize the data into several common chart types by invoking the script language and its official chart library such as the color parsing library, the chart.js file, and

[6]https://github.com/d3/d3/wiki/Gallery
[7]http://www.chartjs.org/

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/ACCESS.2018.2815030, IEEE Access

TABLE 2: Basic information of visualization tools based on JavaScript.

Framework name	Input data format	Rendered charts by	Charts and maps type	License from
D3.js	JSON, CSV, and XML	HTML5 canvas, SVG and CSS	A powerful D3 gallery with multiple charts, graphs, and maps including the world map and the US map.	BSD-3
Chart.js	JavaScript API	Only HTML5 Canvas	8 chart types, including over 23 charts and graphs.	MIT LICENSE
FusionCharts	JSON, XML	SVG, VML	90+ charts and graphs, 1000+ maps including all continents, major countries, and all US states.	Free for basic edition and advanced commercial edition.
Flot Chart	JavaScript API	Only HTML5 Canvas	The charts of lines, points, filled areas, bars and any combinations of these charts. Doesn't support maps.	Free
ZingChart	JavaScript API	HTML5 Canvas, SVG, and VML	Plenty of chart and graph types in its ZingChart gallery. Support almost every country and area.	Free for basic edition and advanced commercial edition.

so on [39]–[41]. Users should insert these libraries into the source code file by coding, then they can use the API from the library to set the parameter and process the chart [42].

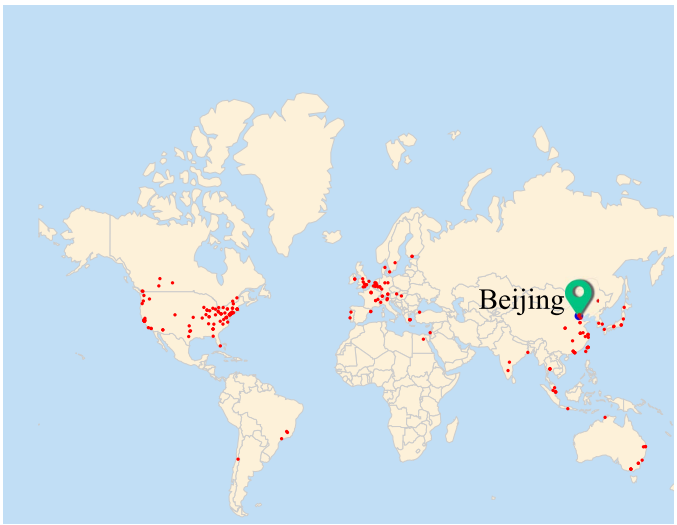

FIGURE 7: The coordinate of different institutions published papers on KDD conference in 2012 using D3.js. Red dots represent the institutions and the green dot represents the conference location [38].

c: FusionCharts

FusionCharts[8] is a commercial JavaScript library suite that combines the technologies including JavaScript and ActionScript3.0. It can run on multi-devices, browsers, and platforms. FusionCharts has more than 90 types of charts and over 1,000 maps which include all of continents. It supports to process .xml and .json files and export the generated charts as .jpg, .png and .pdf files [43]. The function of FusionCharts can be extended to embedding the generated interactive charts to user's applications with several wrappers in official offered plugins, such as JSP charts, PHP charts, jQuery charts, Django charts, etc.

[8]http://www.fusioncharts.com/

d: Flot Charts

Flot Charts[9] focuses on simple usage, attractive exterior and interactive charts. It is an extension of the jQuery library that supports HTML5 charts which is combined canvas and VML. This library separates the functional logic from HTML structure and uses DOM (Document Object Model) element to complete plotting. It contains ready-made components for four basic chart types: charts-bar, line, point, and segment. Users can extend these charts easily and indefinitely with changing a wide variety of configuration parameters [44].

e: ZingChart

ZingChart[10] integrates Angular, React, jQuery, PHP, Ember, and Backbone in its declarative, efficient, and simple JavaScript library. ZingChart supports over 35 types of charts and models and allows users to export their visualization graphs in .png, .jpg, and .pdf formats. It also offers integrated chart arrangement capabilities and has the basic drill-down function that users can select a data item within a chart [45].

2) Tools based on Other Programming Language

a: Gephi

Gephi is a free open-source network visualization software which can implement network analysis. It is written in JAVA on the NetBeans platform. The typical feature of Gephi is that the process of spatialization can be presented vividly. The default layout algorithm of Gephi is ForceAtlas2, it is defined as a continuous force-directed layout algorithm [46]. Users can import their CSV data files or type their data into the spreadsheet of Gephi directly. The data file is divided into two parts: edges table and nodes table, thus users need pre-process their data into two parts. The network supports the number of edges and nodes both up to 1 million, and the visualization is automatical when the data are imported. Users can choose an algorithm (ForceAtlas, ForceAtlas2, Fruchterman Reingold, Noverlap, etc.) to analyze the network, and export the generated network graph in .svg or .pdf formats directly.

[9]http://www.flotcharts.org/
[10]https://www.zingchart.com/

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/ACCESS.2018.2815030, IEEE Access

b: NodeBox

NodeBox[11] is a free open-source and node-based Mac OS X application for creating 2D visuals (static, animated or interactive) that is based on Python programming codes. Users can combine kinds of functionalities optionally by writing Python scripts [47]. NodeBox has integrated various document formats. For example, users can manipulate the vector images in details by invoking the additional SVG library. It also supports the NodeBox Core Image library to create layered images, and exports the generated visuals to a PDF-document and the animations can be exported as QuickTime movies.

c: Ggplot2

Ggplot2 [48] is an open source software package for graphs and visualization of statistical data creating. This package is based on the graphic grammar of R. It allows users to edit the plotting component at a high level of abstraction that is compared with the basic R graphs. This tool attends many details of plotting that makes it fiddly to plot charts or graphs. It is easier to produce complex multi-layer graphics by providing the powerful graphics models and a set of independent building blocks that users can plot a graph piece by piece by implementing the layered grammar of graphics (an extension of Hadley Wickham's grammar of graphics). It means that users will create a more complicated plotting by using faceting that enables users to concentrate more on graphs but less attention to the normalized programming language. Ggplot2 has its wiki in GitGub[12] for providing users an annual case study competition to show their graphs to others in a venue. It also highlights the large range of graphs which are created by using the richness of grammar. Users have the chance to be the developer when they are veteran to ggplot2, as a return, they can contribute codes back to ggplot2.

d: Processing

Processing [49] is an open source programming language based on JAVA that uses the simplified JAVA grammar. It provides users with a graphical interface and runs in the Java environment, and serves as a flexible software sketchbook. Processing is created for teaching fundamentals of computer programming within a visual context. The language has high expansibility that users can write additional codes or integrate existing Java libraries for extending its functions. The official website of Processing[13] is served as the online communication hub to host the relevant references and examples. The website shows a public exhibition about kinds of projects which are designed by Processing [50], [51]. Up to now, Processing provides alternative programming interfaces including JavaScript[14], Python[15], Ruby[16], it also can run on Android for users to create Android application[17].

e: JpGraph

jpGraph[18] is an object-oriented library for creating graphs. The library is based on PHP5 (version above 5.1) and PHP7, and completely written by PHP and compatible with any PHP scripts. The commercial professional version of jpGraph supports the additional graph types: odometer, windrose, and barcodes.

IV. TYPICAL EXAMPLES OF SCHOLARLY ENTITIES VISUALIZATION

Scholarly data contain multiple entities, such as papers, authors, or journals. All these entities can be presented in terms of generated scholarly networks, wherein nodes represent these academic entities and links represent the relationships such as citation, co-author relationship, etc. This section describes the visualization techniques specifically designed for the simple attributes and heterogeneous networks of scholarly data.

A. VISUALIZATION OF ACADEMIC ENTITIES

Scholarly metadata is essential to carry out the efficient management of scholarly documents. Extracting the metadata of a paper such as the title, authors, keywords, algorithms, figures, and tables are vital for developing scholarly services. In order to have a better understanding of topics or trends in science, there are some efforts to visualize the metadata for scholarly documents [52]–[54]. Such efforts are important pieces of scholarly data visualization, with the final goal of expressing how academia develops.

1) Visualization of Researchers

Current bibliographic databases usually provide the service of article searching and author retrieval. Using article metadata constituents such as authors' names, affiliations and research grants to build author profiles can help scientists obtain extensive author-related information [55]. The well-organized information can help scientists to analyze scientific team formation and to have a comprehensive learning about the research interactions in the science of team science.

Name ambiguity, which means many-to-many corresponding relationships between persons and their names [56], is a common problem particularly common in Chinese names in the scientific venues. Shen et al. [54] design a novel visual analytic system for author name disambiguation called NameClarifier. Not same with the traditional black box solution, NameClarifier changes the solution into a white box process with the visual method. Beyond that, it provides a way to guide the users to classify rather than give the

[11] https://www.nodebox.net/code/index.php/Home
[12] https://github.com/tidyverse/ggplot2/wiki
[13] Processing.org
[14] P5.js,https://p5js.org/
[15] Processing.py, http://py.processing.org/
[16] Ruby-Processing, https://github.com/jashkenas/ruby-processing
[17] http://android.processing.org/
[18] http://jpgraph.net/

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/ACCESS.2018.2815030, IEEE Access

TABLE 3: Basic information of visualization tools with other programming language.

Tools	Input data format	Language-based	Features	License from
Gephi	CSV, Excel files,	Java, OpenGL	Powered by OpenGL engine. Force-based layout algorithms shape the graph.	GUN, GPL
Nodebox 3	CSV	Python and Clojure	Integrate all the functional parts in nodes.	GPL
Ggplot2	R, API	R	Plotting based on layers. Graphs composed of layers.	GUN, GPL, and V2
Processing	Multiples of formats are available in its library	Java, plugins for Python and JavaScript	Integrate the OpenGL engine. Over 100+ libraries offered to expand its usage.	GPL, LGPL
JpGraph	CSV, From databases such as mySQL	PHP	Tiny size of generating images. Anti-spam images is supported. 3D effects is also supported.	Free QPL, paid for commercial

classification results simply. The system consists of three parts: Relation View, Temporal View and Group View as well as a list for users to refer back to the original metadata (see FIGURE 8).

In addition, the system also provides rich and practical user interactions, such as view correlation, iterative disambiguation, backtracking, and query. It ensures the effectiveness of person search and sheds light on scientific community detection.

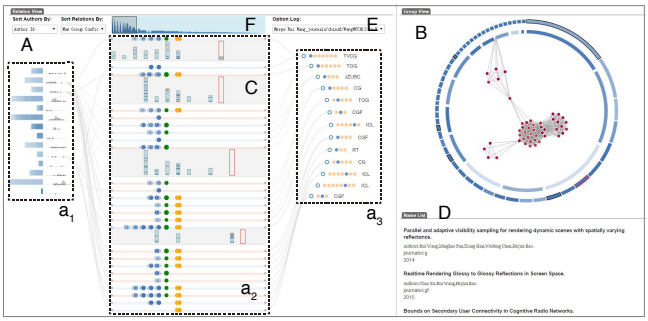

FIGURE 8: Four interfaces for the NameClarifier, which contains: (A) The relation view: contrasts papers which contain ambiguous author names with confirmed authors to classify the ambiguous names easier; (B) The group view: supports the relation view; (C) The temporal view: verifies whether the specific paper can match into a confirmed author's publications; and (D) A list: contains all papers with ambiguous author names, for users to refer back to the original metadata [54].

2) Visualization of Papers

In regards to aggregation levels, papers are the basic research unit, which can be aggregated into several higher research units, such as the author unit, the journal unit, the institution unit, or the country unit. It is an efficient way for researchers to have a comprehensive cognition about their research fields by reading academic papers. By analyzing the scientific literature, it becomes much easier to understand the trends of research or discover links and patterns among scientific documents. It also can help scientists keep track of the latest developments and trends in the hot research topics [57]–[59]. Recently, researchers have shown an increased interest in the visualization of paper to provide multiple views of published articles. Their aim is to discover explicit or implicit relationships between them [60], [61].

Matejka et al. [62] design Citeology System (a portmanteau of citation and genealogy) to explore the relationships between publications which are published at CHI (ACM Conference on Human Factors in Computing Systems) and UIST (ACM Symposium on User Interface Software and Technology). It is implemented as a Java applet and could be a useful tool for finding related work in a specific research field.

Citeology System presents users visualization results on the basis of a "family tree" of sorts, which can represent the generations of the referenced papers built upon on the target paper. Once a paper is selected, it shows the shortest path from hovered paper to the selected paper. Some interesting phenomena could be found through this straightforward visualization of citation generation. For example, it can visualize the main longest direct path between CHI papers which turns out to be an 18 generation gap. Based on these discoveries, it broadens the collection of papers (especially topics) beyond the particular disciplines. The system can make the connection between researchers and the new conferences or topic areas they were not aware of previously.

Portenoy et al. [63] propose a generalizable approach for showing the influence of scholarly fields over time. In this approach, they use a dynamic node-link diagram representing the citation patterns between groups of papers and combine this approach with hierarchical clustering techniques to partition the citation graphs between scholarly papers into clusters which are represented fields and subfields.

There are already numerous tools and techniques for visualizing citation networks, but few focus on the impact of each node in the network. Maguire et al. [64] propose a solution to compare the impact of publications. The design can be divided into three interconnected parts: impact graphs, impact glyphs, and impact overviews (see FIGURE 9). Impact graphs show the specific information of a focus paper, as well as its references and citations. Different patterns are used to identify the varying impact of publications. Impact glyphs are compact versions of the impact graphs to show the comprehensive importance of a paper. Finally, impact overviews position impact graphs for a subject area, author, institution and so on. It represents the core concepts of impact graph and impact glyphs. This part provides a way to

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/ACCESS.2018.2815030, IEEE Access

layout the glyphs in 2D space. The design can outline mass summarizations of publication impact across a database.

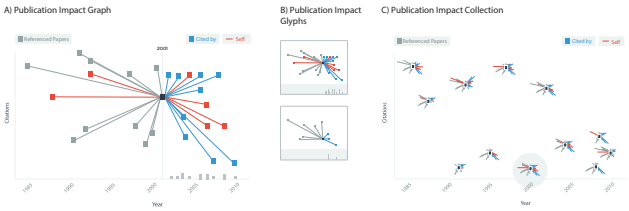

FIGURE 9: An overview of publication impact, which contains: (A) Impact graphs: present the specific information of the paper including references and citations; (B) Impact glyphs: compact the impact graphs to show the comprehensive importance of the paper; (C) Impact overviews: position the impact graphs and impact glyphs for the related information [64].

Jiang et al. [65] conduct the relationships among topics in three different research domains based on a hierarchical topic model. They also provide users a visualization interface and interactive operations to enhance their comprehension of connection among the cross domains, as well as the development trend of visualization. The model aims to represent the hierarchical structure and the similarities between topics. The interactive tool includes five views: word cloud (displays a topic), sankey diagram (represents the evolution of topics), scatter plot (presents the relative position of each topic), treemap (analyzes the relationship of topics), and stream diagram (represents the trends of a topic). FIGURE 10 is the combination of the whole design. It enables users to explore topic mining results interactively and determine the proposed patterns, as well as draw a brief picture of visualization over the past 10 years.

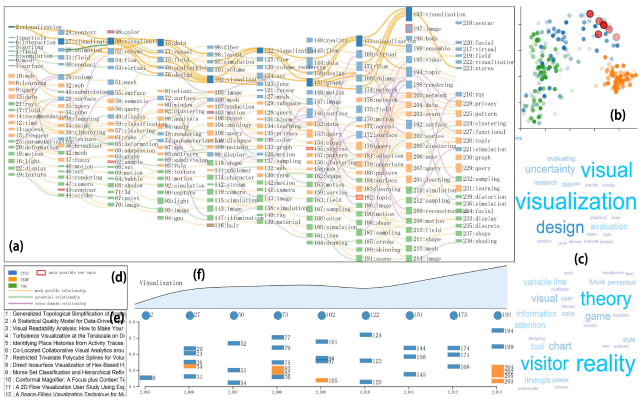

FIGURE 10: An overview of cross-domain-research model, which contains: (a) Sankey diagram: represents the evolution of topics; (b) Scatter plot: shows the topics in 2D space; (c) Word cloud: displays a topic; (d) Legend; (e) Title of selected topics; (f) Stream diagram: represents the trend of a topic [65].

3) Visualization of Institution

Institution is a part of the entities, which makes the scholarly data into a complex system. It encompasses various information including the name, the ranking, members, and location, etc. The relevant information of institutions such as members can be visualized through different techniques (refer to the above). For example, Acemap[19] is a website that can visualize affiliations onto a map (as shown in FIGURE 11). Each node on the map represents an individual affiliation.

For each institution, it can show the collaboration network of the authors. For instance, FIGURE 12 is the collaboration network of Harvard University from the link http://acemap.sjtu.edu.cn/authormap?affID=081E3F30. When clicking on the specific institution in the network, users can also see the total number of publications and authors in the institution easily in the web page. Beyond that, users can choose the specific research field to see the authors in this field.

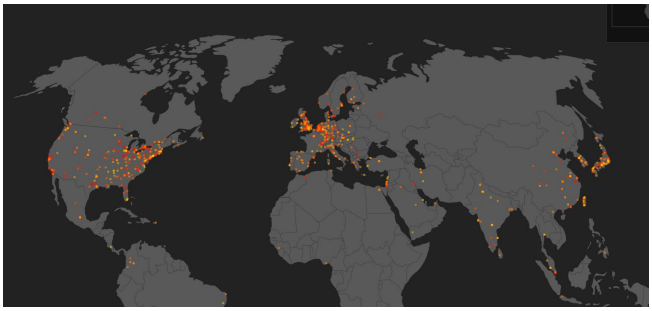

FIGURE 11: An overview of visualizing the institutions contained in the datasets of Acemap. Each node on the map represents an individual institution.

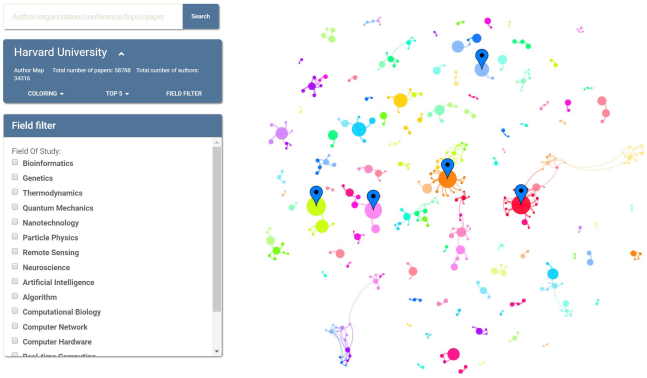

FIGURE 12: The collaboration network of Harvard University in Acemap. Each node represents the author in the institution and the edges between the nodes represent the collaboration between the authors. The color of the nodes represents the research field of the author.

[19]http://acemap.sjtu.edu.cn/app/affiliationMap/index.html

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/ACCESS.2018.2815030, IEEE Access

B. VISUALIZATION OF SCHOLARLY NETWORK

One of the important review articles authored by Newman et al. [66], distinguishes four categories of real-world networks: social networks, information networks, technical networks, and biological networks. Based on this, scholarly networks can be distinguished as social networks (e.g., collaboration networks) and versus information networks (e.g., citation networks).

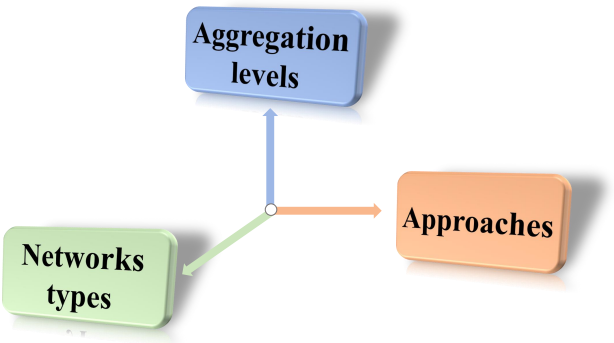

FIGURE 13: Three-dimensional presentation of scholarly network-based bibliometric studies.

1) Visualization of Scholarly Social Network

Recently, there has been renewed interest in the study of various types of scholarly networks [67]. It provides scientists the opportunities to advance the comprehension of the interactive research aggregates [68]. FIGURE 13 presents the three-dimensional presentation of scholarly network-based bibliometric studies including approaches, networks types, and aggregation levels. Social network analysis can be divided into two typical ways: personal center network analysis (ego-centric analysis), and group network analysis (sociocentric analysis) [69]. On this basis, we show the current conditions of study in scholarly networks visualization from following two aspects: visualization of scholarly ego-centric network and visualization of scholarly sociocentric network.

Ego is the central node of the ego-centric network, and alters are associated nodes. Depending on the distance from the alters to the ego, alters can be divided into 1-degree alter (nodes connected with ego directly), 2-degree alter (nodes connected with ego's alters), and so on. The ego-centric network focuses on the impact of the network on the ego. The networks have multidimensional attributes and change with time. How to display these characteristics that can make scientists understand, analyze and solve practical problems becomes a central issue of visual research.

As shown in FIGURE 14, the scholarly tree is designed by Fung et al. [13] based on a botanic tree metaphor. It is a web-based, interactive visual interface. The different parts of the tree (e.g., leaves, branches and trunks) are on behalf of different characteristics of the scholars' published papers. This project aims to show the details of collaboration information based on the unique tree features. The patterns of visualization encourage scientists to examine the personal career and also help to promote their self-development.

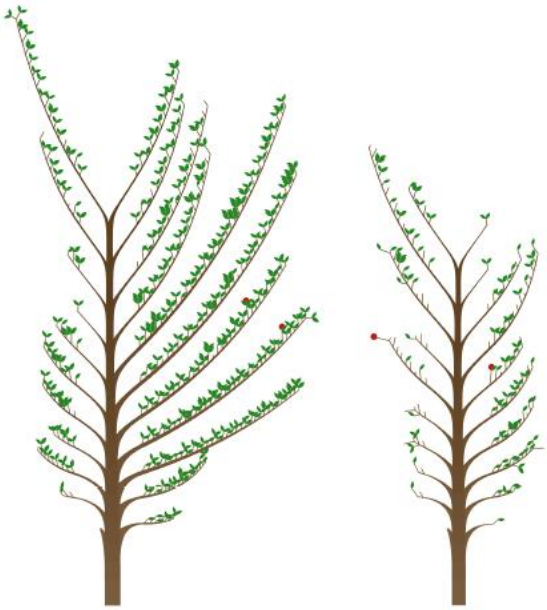

FIGURE 14: The scholarly tree of two active researchers. Each branch on the tree encodes the publications of two years. The trees display their details of the publications between 1993-2013 for the left one and 1995-2012 for the right one [13].

Botanical tree is focused on the association between the ego and 1-degree alters and emphasizes on the information representation. In order to figure out the evolution of personal networks from different perspectives, Wu et al. [70] design the egoSlider System. FIGURE 15 uses different colors and shapes on the ring to show the relationship between the ego and the alters. It also presents the number of 1-degree alters (stripes) sustaining in a continuous-time period based on the glyph encoding. The glyph focuses on showing the overall statistical information, which is different from the line chart-based exhibition of the specific information about 1-degree alters and 2-degree alters. The design provides a wealth of interaction. It can change the glyph coding and the group method of lines flexibly according to user requirements.

EgoSlider can be applied to explore the DBLP collaboration network. It can help to discover a scientist's academic career, the change of the research direction, and the closeness of the collaboration with other scientists. Moreover, it also can make a major contribution for explaining some interesting phenomena by combing with the practical information found in the system. Compared with the baseline system based on the node link diagram, it will be more efficient to use egoSlider to complete the same task.

2) Visualization of Scholarly Information Network

Co-citation count represents times of two papers cited jointly. Visualization of co-citation contributes much to the under-

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/ACCESS.2018.2815030, IEEE Access

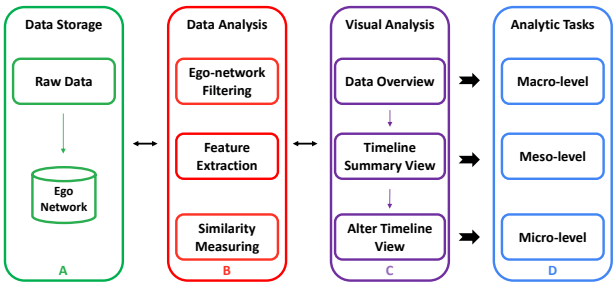

FIGURE 15: The overview of the egoSlider visualization pipeline, which contains four parts: (A) Data storage: extracts the ego-network structures from the raw data and stores into MongoDB; (B) Data analysis: integrates several analytical methods to process the dynamic ego-network sequences; (C) Visual analysis: performs visual analysis of the data to let users interactively navigate; (D) Analytic tasks: addresses a different level of ego-centric analytical tasks.

standing of similarity in the document influence network. Early efforts in visualizing document co-citation similarity employed cluster visualization. Noel et al. [71] provide examples of influence network visualization for both co-citation count and co-citation correlation and observe that the correlation-based visualization exhibits chaining effects.

Recently, Wu et al. [72] have designed PathWay to discover and understand the trends in the bibliographic data on the basis of individual professionals' co-authorship as well as the citation network of their publications. Implemented with JavaScript, HTLM5, and the Scalable Vector Graphics (SVG), the system displays researchers' career paths in terms of their collaboration networks. The design can help users understand the social process better under the situation of challenges emerging.

V. EXISTING VISUAL ANALYTIC SYSTEMS FOR SCHOLARLY DATA

For conducting the research of scholarly data, researchers usually need to extract the basic information of the large-scale scholarly datasets (such as DBLP, WOS, MAG) for analyzing the science of science. There are some visual analytic systems for scholarly data, which has brought a great convenience for researchers. It costs less time for users to concentrate on processing and analyzing the scholarly data by using these systems. In this section, we introduce 5 typical visual analytic systems for scholarly data below. In TABLE 4, we list features for comparing the basic information of these five visual analytic systems.

A. VOSVIEWER

VOSviewer is a free visualization and analysis tool for constructing and visualizing bibliometric networks[20]. It facilitates the analysis of clustering solutions by visualizing the scholarly data into bibliometric networks. Actually, another tool named CitNetExplorer[21] is also used to cluster publications and analyze the resulting clustering solutions. CitNetExplorer focuses on the analysis at an individual publications level, while VOSviewer focuses on the analysis at an aggregate level. Users can create the network by importing the data files from Web of Science (WOS), Scopus, PubMed[22], Reference Information Systems (RIS), Pajek, and Graph Modelling Language (GML). VOSviewer can create biblometric networks and handle them by using the advanced layouts and clustering techniques. The visualization of these networks can be saved as a bitmap file or in vector format [73]. Xu et al. [74] present a bibliographic coupling analysis of the 20 most productive countries/territories in TFS (IEEE Transactions on Fuzzy Systems) publications based on VOSviewer (as shown in FIGURE 16).

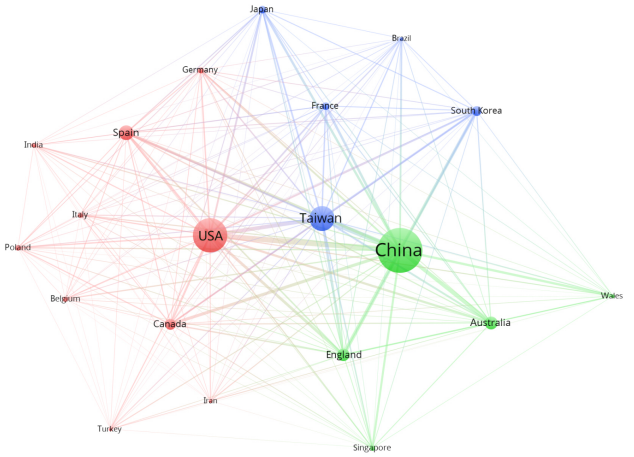

FIGURE 16: The bibliographic coupling of the 20 most productive countries/regions in TFS publications in VOSviewer. The biggest circle represents China, which means that China is the most productive country, and the second is the United States [74].

B. SCI2

The Science of Science (Sci2 [75]) tool[23] is an open source modular toolset which supports the temporal, geospatial, topical, and network studies. Sci2 can visualize the scholarly datasets into different kinds of networks. The visualization of small datasets can be explored interactively, and the large-scale datasets are rendered into Postscript files that users can convert it into .pdf files and retrieve its information as a filter, such as searching the specific text in the visualization [14]. Osili et al. [76] geolocate donors and recipients based on their combined cities, states, and country information in existing data using the Bing! geocoder[24] available in the

[20] VOSviewer,http://www.vosviewer.com/

[21] CitNetExplorer,http://www.citnetexplorer.nl/

[22] https://www.ncbi.nlm.nih.gov/pubmed

[23] https://sci2.cns.iu.edu/user/index.php

[24] http://wiki.cns.iu.edu/display/CISHELL/Bing+Geocoder

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/ACCESS.2018.2815030, IEEE Access

TABLE 4: Basic information of visual analytic systems for scholarly data.

System name	Main function	Supported data file format	Features	Operating environment	Limitation and the key difference
VOSViewer	Taking citation analysis.	Data from WOS, Scopus, PubMed, RIS, Pajek, and GML.	Density and overlay visualizations. Create bibliometric networks based on co-authorship, bibliographic coupling, and co-citation networks, etc. Natural language processing techniques are available for creating term co-occurrence networks.	Windows, Mac OS X, Other systems with the support of Java 6 or later updates, the web client based on Java installed.	Only support node network diagram and corresponding heat map. Strong graphic display ability, suitable for large-scale scholarly data.
Sci2	Taking network analysis, especially directed network analysis.	TXT, CSV, Network data (in-memory graph/network object or network files saved as Graph/ML, XGMML, NWB, Pajek .net or Edge list format), Matrix data (Pajek.mat), In-memory database, Tree data (TreeML)	Visualize the scholarly datasets into kinds of networks. Perform different types of analysis with the most effective algorithms available. Access science datasets online or load their own.	Mac OS, Windows, Linux	High memory footprint while processing large datasets. Temporal bar graph, Choropleth map, UCSD science map and Bimodal network visualization are supported.
Histcite	Taking statistic analysis.	Data from WOS	Scholarly data visualization and various types of bibliometric analysis	Windows, based on the IE browser	No longer in official supported. Only support WOS data. Diagrams the development of one scientific field based on the timeline.
Bibexcel	Processing scholarly database.	Plain text data from WOS, SCIE, DII (Derwent Innovations Index), Medline	Able to do various types of bibliometric analysis. Export its processed data into other visualization tools (Gephi, Pajek, VOSviewer, etc.) that can take a comprehensive visual analysis.	Windows	Flexible in processing scholarly data but not easy to use without its help document.
CiteSpace	Taking co-citation analysis.	Data from WOS, Scopus, PubMed, ADS, arXiv, CNKI, CSSCI, Derwent, NSF, Project DX, ADS, and CiteSpace Built-in database.	Visualize and analyze the patterns and trends in scientific literature	Windows, Require Java 8	Can not delete the irrelevant node within the generated node network.

Sci2 tool. They also extract a bi-modal network of the major donors and the six merged sub-sectors and visualize them by using the Sci2 tool. The detailed instructions are available in http://wiki.cns.iu.edu/display/SCI2TUTORIAL/Bipartite+Network+Graph.

C. HISTCITE

Histcite [78] is a software package that runs on Windows system with the Internet Explorer. This system is used for scholarly data visualization and bibliometric analysis including the productive authors, the scale of journals, the frequency of words, the type of documents, and the ranking of countries/institutions. Histcite converts the bibliographies datasets into time-based networks called historiograph, and makes it easier for users to observe and understand the main publishing events of the subject and the impact of the chronology in the networks [79].

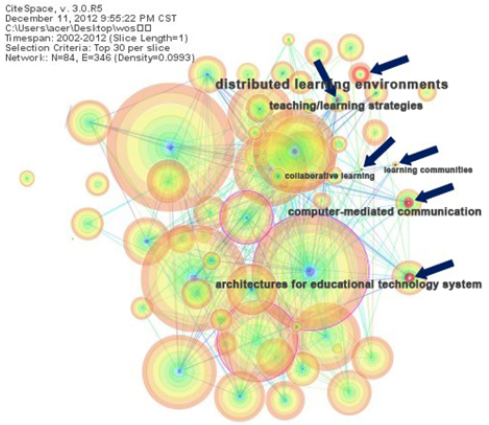

FIGURE 17: The visualized network in CiteSpace which shows the co-cited references and burst terms of e-learning research (2002-2013) [77].

D. BIBEXCEL

BibExcel is a multi-functional bibliometric toolbox developed by Persson [80]. BibExcel is used to do various types of bibliometric analysis, such as citation analysis, cluster analysis, co-citation analysis, and so on. It allows users to analyze their scholarly data by selecting a catalogue which exists in their data (such as the authors) and adding it as a variable in a data matrix of the output files which is created by BibExcel. This software also allows users to export the files which include the data matrix and can import to other visualization tools (Gephi, Pajek, and VOSviewer, etc.) for visualization [81].

E. CITESPACE

Citespace[25] is a free Java application developed by Chen [21], [82]. It runs on the java virtual machine so it requires the Java runtime environment. The goal of this software is to detect, visualize and analyze emerging trends and critical changes in the scientific literature. It combines information visualization methods, bibliometrics with the algorithms of data mining in an interactive visualization tool to extract the patterns in citation data. Lin et al. [77] visualize the co-cited references and burst terms of e-learning research into a network by using CiteSpace (as shown in FIGURE 17).

VI. OPEN ISSUES AND OUTLOOK

The majority of the techniques and systems discussed in this survey specifically address one or two facets of scholarly data visualization. Benefiting from the development and popularization of these techniques, scientists have opportunities to study "science of science" from a new perspective. However, the rapid growing scholarly data also brings numerous challenges to the field of scholarly data visualization. We discuss the issues that seem promising for further research as follows.

A. INFORMATION INTEGRATION

One of the main challenges in scholarly data visualization is information integration. Previous visualization tasks mainly focus on presenting the relationships extracted from publications, for example, citation visualization or collaboration visualization. However, scholarly data contain various entities including papers, authors, institutions, etc. There are diverse relationships among these entities. How to visualize different relationships in a single task is meaningful and challenging. At the same time, scholarly relationships are hidden in different data sources. For example, the collaboration information can be gained from the digital libraries, i.e., DBLP, while the friend relationships are hidden in online social networks, i.e., Facebook. How to integrate the information from different data sources is a promising open issue in scholarly data visualization.

B. VISUALIZATION TECHNIQUES

Although a number of visualization methods have been proposed, some specific scholarly visualization techniques are encouraged to be improved. For example, very little attention has been paid to the visualization of the academic institutions. Another problem is that scholarly data often contain abundant information. How to mine useful information with the help of visualization tools is still a critical problem to be solved. The effectiveness needs to be enhanced due to the increasing complex network structure as well.

Another challenge is how to combine the visualization techniques with the scholarly data analysis efficiently. The visualization theories and techniques on scholarly data visual analysis are not well integrated in practice. The ability of data processing of these visualization techniques also needs to be improved. Generally, researchers acquire scholarly datasets from different online scholarly data platforms. The dataset may be very huge and researchers need to pre-process these heterogeneous data (data merging, data partition, delete unrecognized attributes in the dataset, etc.) for meeting the data-input requirements of these visualization techniques. How to enhance data processing ability of visualization techniques is also important and promising.

C. VISUALIZATION ANALYSIS

In the process of analyzing the scholarly visualization results, there is still no specific benchmark to evaluate the merits between visualization schemes. For instance, for evaluating the international impact of a given scholar, there are two visualization schemes based on the same database. One is the map visualization of the country distribution in co-citation relationship of the given scholar. Another is the map visualization of the country distribution in co-author relationship. It is hard for us to analyze the merits of the above visualization results because both co-citation relationship and co-author relationship can represent scholars' impact. But which attribute is the stronger one to show the impact of the scholar? Which attribute can effectively and accurately reflect one's scholarly impact? We need to develop the standard rubrics and benchmarks for evaluating the merits of visualization schemes.

D. VISUALIZATION ANALYTIC SYSTEMS

The most typical limitation of the existing visualization analytic systems is the pertinency of the dataset. Most of these systems support only few types of database files such as WOS, Scopus and PubMed files. Thus, these visualization analytic systems cannot serve to analyze the database with other types. Another problem in existing visualization analytic systems is the ability of data processing. The increasing size of the data will take up a lot of solid-state memory. Ordinary computers cannot meet the operational needs. How to optimize these systems to enhance their usability is also an open issue.

[25]Citespace,https://sites.google.com/site/citespace101/

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/ACCESS.2018.2815030, IEEE Access

VII. CONCLUSION

Scholarly big data bring a variety of opportunities and challenges of scholarly data analysis. Nowadays, researchers have realized the significance of applying the visualization technologies on different datasets to comprehend the science itself. Thus scholarly data visualization plays a key role in addressing the problems arising from large-volume, multi-variety, and important-value data. It makes sense to concentrate more on this topic.

To provide new insights into scholarly data visualization, we review the emerging area of it in this survey paper. We present state-of-the-art scholarly data visualization techniques with a focus on the visualization tools and analytic systems. According to the characteristics of scholarly data, visualization techniques for scholarly data are presented in two aspects: the single attribute and heterogeneous networks. Meanwhile, various scholarly data analytic systems are developed to compile visual analysis of multivariate data from various aspects (e.g., citation relationship, co-citation relationship, and co-authorship). Therefore, this study makes a major contribution to research on scholarly data visualization by demonstrating details of the topic. Although it becomes the focus of the current research, some technologies still need to be improved. One of the major challenges is how to integrate information from the complex scholarly data efficiently. Another challenge is how to combine various visualization techniques with the analysis processing more suitably. A future study investigating these questions would be very significant.

REFERENCES

[1] 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.

[2] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," IEEE Transactions on Big Data, vol. 3, no. 1, pp. 18–35, 2017.

[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," in Proceedings of the Workshop: Web-Scale Classification: Classifying Big Data from the Web, New York, NY, 2014.

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

[5] 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 Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on. IEEE, 2016, pp. 209–210.

[6] M. E. M. Barak and J. S. Brekke, "Social work science and identity formation for doctoral scholars within intellectual communities," Research on Social Work Practice, vol. 24, no. 5, pp. 616–624, 2014.

[7] G. Cormode, S. Muthukrishnan, and J. Yan, "People like us: Mining scholarly data for comparable researchers," in Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014, pp. 1227–1232.

[8] C. D. Hansen and C. R. Johnson, Visualization handbook. Academic Press, 2011.

[9] T. L. Naps, G. Rößling, V. Almstrum, W. Dann, R. Fleischer, C. Hundhausen, A. Korhonen, L. Malmi, M. McNally, S. Rodger et al., "Exploring the role of visualization and engagement in computer science education," in ACM Sigcse Bulletin, vol. 35, no. 2. ACM, 2002, pp. 131–152.

[10] F. Xia, X. Su, W. Wang, C. Zhang, Z. Ning, and I. Lee, "Bibliographic analysis of nature based on twitter and facebook altmetrics data," PloS one, vol. 11, no. 12, p. e0165997, 2016.

[11] B. H. McCormick, T. A. DeFanti, and M. D. Brown, "Visualization in scientific computing," IEEE Computer Graphics and Applications, vol. 7, no. 10, pp. 69–69, 1987.

[12] D. A. Keim, "Information visualization and visual data mining," IEEE transactions on Visualization and Computer Graphics, vol. 8, no. 1, pp. 1–8, 2002.

[13] T. L. Fung and K.-L. Ma, "Visual characterization of personal bibliographic data using a botanical tree design," in Proceedings of IEEE VIS 2015 Workshop on Personal Visualization: Exploring Data in Everyday Life, 2015.

[14] 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.

[15] I. Lee, F. Xia, and G. Roos, "An observation of research complexity in top universities based on research publications," in Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 1259–1265.

[16] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, Mastering the information age solving problems with visual analytics. Eurographics Association, 2010.

[17] X. Bai, J. Hou, H. Du, X. Kong, and F. Xia, "Evaluating the impact of articles with geographical distances between institutions," in Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 1243–1244.

[18] J. Li, C. Liu, L. Chen, Z. He, A. Datta, and F. Xia, "iTopic: Influential topic discovery from information networks via keyword query," in Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 231–235.

[19] C. Chen, "Visualising semantic spaces and author co-citation networks in digital libraries," Information processing & management, vol. 35, no. 3, pp. 401–420, 1999.

[20] W. Wang, J. Liu, F. Xia, I. King, and H. Tong, "Shifu: Deep learning based advisor-advisee relationship mining in scholarly big data," in Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 303–310.

[21] C. Chen, "Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature," Journal of the American Society for information Science and Technology, vol. 57, no. 3, pp. 359–377, 2006.

[22] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in Proceedings of the 24th international conference on world wide web. ACM, 2015, pp. 243–246.

[23] 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.

[24] W. Wang, S. Yu, T. M. Bekele, X. Kong, and F. Xia, "Scientific collaboration patterns vary with scholars' academic ages," Scientometrics, vol. 112, no. 1, pp. 329–343, 2017.

[25] H. Li, I. Councill, W.-C. Lee, and C. L. Giles, "CiteSeerx: An architecture and web service design for an academic document search engine," in Proceedings of the 15th international conference on World Wide Web. ACM, 2006, pp. 883–884.

[26] L. Yao, J. Tang, and J. Li, "A unified approach to researcher profiling," in Web Intelligence, IEEE/WIC/ACM International Conference on. IEEE, 2007, pp. 359–366.

[27] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 990–998.

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

[29] E. Cortez, A. S. da Silva, M. A. Gonçalves, F. Mesquita, and E. S. de Moura, "Flux-cim: Flexible unsupervised extraction of citation metadata," in Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries. ACM, 2007, pp. 215–224.

[30] F. Peng and A. McCallum, "Information extraction from research papers using conditional random fields," Information processing & management, vol. 42, no. 4, pp. 963–979, 2006.

[31] 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.

[32] 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.

[33] J. Hagerty, R. L. Sallam, and J. Richardson, "Magic quadrant for business intelligence platforms," Gartner for Business Leaders (February 6, 2012), 2012.

[34] P. Kale and S. Balan, "Big data application in job trend analysis," in Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016, pp. 4001–4003.

[35] X. Bai, H. Liu, F. Zhang, Z. Ning, X. Kong, I. Lee, and F. Xia, "An overview on evaluating and predicting scholarly article impact," Information, vol. 8, no. 3, p. 73, 2017.

[36] M. Bostock, "D3. js-data-driven documents," URL: https://d3js. org, 2016.

[37] F. Bao and J. Chen, "Visual framework for big data in d3.js," in Electronics, Computer and Applications, 2014 IEEE Workshop on. IEEE, 2014, pp. 47–50.

[38] X. Bai, F. Zhang, J. Hou, F. Xia, A. Tolba, and E. Elashkar, "Implicit multi-feature learning for dynamic time series prediction of the impact of institutions," IEEE Access, vol. 5, pp. 16 372–16 382, 2017.

[39] N. Downie, "Chart.js: Open source html5 charts for your website," Chart.js, 2015.

[40] C. Bergstrom, "Eigenfactor: Measuring the value and prestige of scholarly journals," College & Research Libraries News, vol. 68, no. 5, pp. 314–316, 2007.

[41] R. Murphy, "An employee performance simulation to aide in managerial decision making in a target driven work environment," 2016.

[42] R. Raghav, S. Pothula, T. Vengattaraman, and D. Ponnurangam, "A survey of data visualization tools for analyzing large volume of data in big data platform," in Communication and Electronics Systems (ICCES), International Conference on. IEEE, 2016, pp. 1–6.

[43] S. Nadhani and P. Nadhani, FusionCharts beginner's guide: The official guide for FusionCharts suite. Packt Publishing Ltd, 2012.

[44] P. Pokornỳ and K. Stokláska, "Chart visualization of large data amount," in Computer Science On-line Conference. Springer, 2017, pp. 460–468.

[45] R. L. Rothfeld, "Advancing web-based dashboards: Providing contextualised comparisons in an air traffic discovery dashboard," 2015.

[46] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," PloS one, vol. 9, no. 6, p. e98679, 2014.

[47] T. De Smedt, L. Lechat, and W. Daelemans, "Generative art inspired by nature, using nodebox," Applications of Evolutionary Computation, pp. 264–272, 2011.

[48] H. Wickham, ggplot2: elegant graphics for data analysis. Springer, 2016.

[49] C. Reas and B. Fry, "Processing.org," Processing.org, vol. 3, no. 06, 2012.

[50] A. Bigelow, S. Drucker, D. Fisher, and M. Meyer, "Reflections on how designers design with data," in Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces. ACM, 2014, pp. 17–24.

[51] C. Yang, I. Jensen, and P. Rosen, "A multiscale approach to network event identification using geolocated twitter data," Computing, vol. 96, no. 1, pp. 3–13, 2014.

[52] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, "Visualization as seen through its research paper keywords," IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 1, pp. 771–780, 2017.

[53] J. Whittaker, "Creativity and conformity in science: Titles, keywords and co-word analysis," Social Studies of Science, vol. 19, no. 3, pp. 473–496, 1989.

[54] Q. Shen, T. Wu, H. Yang, Y. Wu, H. Qu, and W. Cui, "Nameclarifier: A visual analytics system for author name disambiguation," IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 1, pp. 141–150, 2017.

[55] J. Zhang, Z. Ning, X. Bai, X. Kong, J. Zhou, and F. Xia, "Exploring time factors in measuring the scientific impact of scholars," Scientometrics, vol. 112, no. 3, pp. 1301–1321, 2017.

[56] J. Kim, H. Kim, and J. Diesner, "The impact of name ambiguity on properties of coauthorship networks," Journal of Information Science Theory and Practice, vol. 2, no. 2, pp. 6–15, 2014.

[57] M. Callon, J. P. Courtial, and F. Laville, "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry," Scientometrics, vol. 22, no. 1, pp. 155–205, 1991.

[58] S. Ravikumar, A. Agrahari, and S. Singh, "Mapping the intellectual structure of scientometrics: A co-word analysis of the journal scientometrics (2005–2010)," Scientometrics, vol. 102, no. 1, pp. 929–955, 2015.

[59] J. Law, S. Bauin, J. Courtial, and J. Whittaker, "Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification," Scientometrics, vol. 14, no. 3-4, pp. 251–264, 1988.

[60] A. Khan, J. Matejka, G. Fitzmaurice, and G. Kurtenbach, "Spotlight: Directing users' attention on large displays," in Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2005, pp. 791–798.

[61] F. Heimerl, Q. Han, S. Koch, and T. Ertl, "Citerivers: Visual analytics of citation patterns," IEEE transactions on visualization and computer graphics, vol. 22, no. 1, pp. 190–199, 2016.

[62] J. Matejka, T. Grossman, and G. Fitzmaurice, "Citeology: Visualizing paper genealogy," in CHI'12 Extended Abstracts on Human Factors in Computing Systems. ACM, 2012, pp. 181–190.

[63] J. Portenoy and J. D. West, "Dynamic visualization of citation networks showing the influence of scholarly fields over time," in International Workshop on Semantic, Analytics, Visualization. Springer, 2016, pp. 147–151.

[64] E. Maguire, J. M. Montull, and G. Louppe, "Visualization of publication impact," in Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers. Eurographics Association, 2016, pp. 103–107.

[65] X. Jiang and J. Zhang, "A text visualization method for cross-domain research topic mining," Journal of Visualization, vol. 19, no. 3, pp. 561–576, 2016.

[66] M. E. Newman, "The structure and function of complex networks," SIAM review, vol. 45, no. 2, pp. 167–256, 2003.

[67] J. Zhang, F. Xia, Z. Ning, T. M. Bekele, X. Bai, X. Su, and J. Wang, "A hybrid mechanism for innovation diffusion in social networks," IEEE Access, vol. 4, pp. 5408–5416, 2016.

[68] 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.

[69] J. Scott, Social network analysis. Sage, 2017.

[70] Y. Wu, N. Pitipornvivat, J. Zhao, S. Yang, G. Huang, and H. Qu, "Egoslider: Visual analysis of egocentric network evolution," IEEE transactions on visualization and computer graphics, vol. 22, no. 1, pp. 260–269, 2016.

[71] S. Noel, C.-H. H. Chu, and V. Raghavan, "Co-citation count vs correlation for influence network visualization," Information Visualization, vol. 2, no. 3, pp. 160–170, 2003.

[72] M. Q. Y. Wu, R. Faris, and K.-L. Ma, "Visual exploration of academic career paths," in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM, 2013, pp. 779–786.

[73] N. J. van Eck and L. Waltman, "Citation-based clustering of publications using citnetexplorer and vosviewer," Scientometrics, vol. 111, no. 2, pp. 1053–1070, 2017.

[74] Z. Xu, D. Yu, Y. Kao, and C.-T. Lin, "The structure and citation landscape of ieee transactions on fuzzy systems (1994-2015)," IEEE Transactions on Fuzzy Systems, 2017.

[75] S. Team, "Science of science (sci2) tool," Indiana University and SciTech Strategies, 2009.

[76] U. O. Osili, J. Ackerman, C. H. Kong, R. P. Light, and K. Börner, "Philanthro-metrics: Mining multi-million-dollar gifts," PloS one, vol. 12, no. 5, p. e0176738, 2017.

[77] X.-F. Lin and Q. Hu, "Trends in e-learning research from 2002-2013: A co-citation analysis," in Advanced Learning Technologies (ICALT), 2015 IEEE 15th International Conference on. IEEE, 2015, pp. 483–485.

[78] T. Reuters, "Histcite," http://interest.science.thomsonreuters.com/forms/HistCite/, 2014, [accessed 29-september-2017].

[79] E. Garfield, "From the science of science to scientometrics visualizing the history of science with histcite software," Journal of Informetrics, vol. 3, no. 3, pp. 173–179, 2009.

[80] O. Persson, "Bibexcel: A toolbox for bibliometricians," Tech. Rep.

[81] O. Persson, R. Danell, and J. W. Schneider, "How to use bibexcel for various types of bibliometric analysis," Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday, vol. 5, pp. 9–24, 2009.

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/ACCESS.2018.2815030, IEEE Access

[82] C. Chen, "The citespace manual," 2014.

JIAYING LIU received the BS degree in software engineering from Dalian University of Technology, China, in 2016. She is currently working toward the master's degree in the School of Software, Dalian University of Technology, China. Her research interests include big scholarly data, social network analysis, and science of success.

TAO TANG is an undergraduate in Chengdu College, University of Electronic Science and Technology of China, Chengdu, China. He is currently working toward the Bachelor's degree in computer science and technology. His research interests include big data analytics and visualization.

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 in Dalian University of Technology (DUT), Dalian, China. His research interests include big scholarly data, social network analysis, and computational social science.

BO XU received the BSc and PhD degrees from the Dalian University of Technology, China, in 2007 and 2014, respectively. She is currently a lecturer in School of Software at the Dalian University of Technology. Her current research interests include social computing, data mining, information retrieval, and natural language processing.

XIANGJIE KONG (M'13-SM'17) received the BSc and PhD degrees from Zhejiang University, Hangzhou, China. He is currently an Associate Professor in School of Software, Dalian University of Technology, China. He has served as (Guest) Editor of several international journals, Workshop Chair or PC Member of a number of conferences. Dr. Kong has published over 50 scientific papers in international journals and conferences (with 30+ indexed by ISI SCIE). His research interests include intelligent transportation systems, mobile computing, and cyber-physical systems. He is a Senior Member of IEEE and CCF, and a Member of ACM.

FENG XIA (M'07-SM'12) received the BSc and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently a Full Professor in School of Software, Dalian University of Technology, China. Dr. Xia has published 2 books and over 200 scientific papers in international journals and conferences. His research interests include computational social science, network science, data science, and mobile social networks. He is a Senior Member of IEEE and ACM.