



Min Tian · Guozheng Li · Xiaoru Yuan

LitVis: a visual analytics approach for managing and exploring literature

Received: 11 July 2021 / Revised: 28 January 2023 / Accepted: 30 July 2023 / Published online: 28 September 2023
© The Visualization Society of Japan 2023

Abstract Reading literature is essential to research. However, the explosive growth, the multidimensional attributes, and the complex relationships pose a tremendous challenge for researchers to understand and analyze literature efficiently. We propose LitVis, a visual analysis approach to help users manage and explore literature based on its metadata. LitVis allows users to select literature collection of interest and analyze them from their attributes, text, and citation networks. From the perspective of attribute values, LitVis supports users in understanding the distribution of literature and filtering individuals of interest. From the perspective of the text, LitVis uses the Latent Dirichlet Allocation model to extract topics from the literature and allows users to adjust the topic extraction results interactively. From the citation network perspective, LitVis enables users to analyze citation relationships within and between topics to help them understand research development. One use case and carefully designed interviews with domain experts validate the effectiveness of LitVis in the management and analysis of the literature. The results show that LitVis help users comprehensively identify the literature collection of interest and efficiently analyze the evolution of research topics.

Keywords Literature visualization · Graph/network visualization · Text visualization

1 Introduction

The research literature is the embody of the scientific knowledge (Fortunato et al. 2018). During the process of research, researchers publish their exploration results into the literature at conferences and journals. As an essential output that reflects the development and progress of research, literature documents the evolution and trends of research studies, the interactions between research topics, and the collaborative relationships

M. Tian · X. Yuan
Key Laboratory of Machine Perception (Ministry of Education), Beijing, China
E-mail: tianmin@pku.edu.cn

M. Tian · X. Yuan
School of AI, Peking University, Beijing, China

G. Li
School of Computer Science, Beijing Institute of Technology, Beijing, China
E-mail: guozheng.li@bit.edu.cn

X. Yuan (✉)
National Engineering Laboratory for Big Data Analysis and Application, Peking University, Beijing, China
E-mail: xiaoru.yuan@pku.edu.cn

between researchers. Managing and analyzing literature are a critical task in both summarizing existing work and conducting further research.

Data availability prompts a great deal of research work on literature visualization. Diverse data sources about research literature are available today, including Web of Science(WoS),¹ Google Scholar,² Microsoft Academic,³ covering millions of literature. Some of these data sources are even freely accessible. To understand the complex and abstract literature dataset, researchers develop various visualizations in many usage scenarios of literature analysis, such as identifying the evolution (Dou et al. 2013; Lee et al. 2005; Liu et al. 2012) and emergence (Small et al. 2014) of research topics, creating collaboration and citation networks to reveal the skeleton and composition (Klavans and Boyack 2009). Existing studies either focus on the visualization of a large-scale literature dataset or a single publication. However, existing methods do not support users to select the collection of interest and analyze the inter-relationships and evolution at the topic level. One typical application scenario is to analyze the research work in a particular research field and write the literature review, e.g., research work about social media visualization and graph visualization. These studies may vary significantly in research questions but are closely related to each other from methodology aspects. For example, some studies may utilize a general method to various aspects along a line of thought, and some studies may apply the general methods to different application scenarios. In such a case, the objective of users' exploration is the literature collection related to social media. The task is to extract topics, understand their evolution, and analyze the correlations and intersections between various topics.

We propose LitVis, which provides a novel visual analytics perspective for managing and exploring literature dataset. Firstly, it allows users to determine a literature collection of interest interactively. After loading the literature dataset, LitVis visualizes the publications using unit visualization (Stasko et al. 2013) and arranges the literature according to the publication date and venues. To select the literature collection of interest, LitVis allows users to filter the publications according to the metadata attributes, citation relationships, and the tags and comments added on the publications by users. Furthermore, LitVis extracts topics according to the textual data of literature through the Latent Dirichlet Allocation technique (LDA) (Blei et al. 2003) and allows users to adjust the extracted topic interactively. Based on the topic modeling results, LitVis aggregates the citation relationships between publications at the topic level and enables users to analyze the relationships among topics at different granularities. By arranging the publications chronologically, users can also understand the evolution of research topics.

To evaluate the usability and effectiveness of LitVis, we had 10 graduate students use the prototype system to analyze publications collections related to one specific research field. We report one case study to demonstrate the utility of LitVis. Results show that LitVis can facilitate users to manage publications efficiently and build their comprehensive understanding of the inter-relationship and evolution among different research topics of the literature dataset.

In conclusion, our contributions can be summarized as:

- A novel analytical perspective for visualizing the inter-relationships and evolution of literature at the topic level;
- A visual analytic system, LitVis, to support users manage, filter, and explore research topics of a set of interested publications;
- One use case involving the publication collections of one research topic to demonstrate the usability and effectiveness of LitVis.

In the remainder of this paper, we review the background with related work in Sect. 2. In Sect. 3, we discuss the design requirements of the literature visualization. In Sect. 4, we introduce the system design of LitVis, including interface design and interaction design. We provide a case study for analyzing the publication collection of one research field in Sect. 5 to demonstrate the effectiveness of LitVis, followed by the discussions in Sect. 6 about the method comparison and future work. In Sect. 7, we present the conclusions of our paper.

¹ <https://www.webofknowledge.com/>.

² <https://scholar.google.com/>.

³ <https://academic.microsoft.com/>.

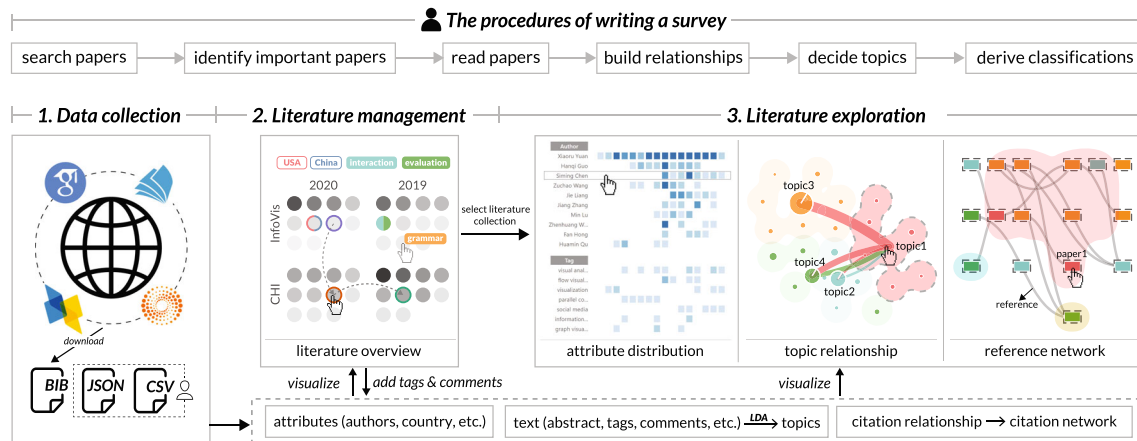


Fig. 1 The top part shows the summarized procedures of writing a survey. The bottom part shows the pipeline of LitVis, composing three modules. (1) Data collection requires users to download the bib files from the literature databases (e.g., google scholar) or prepare the literature dataset in JSON/CSV format manually; (2) Literature management facilitates users to understand the information of literature and allows them to add tags or comments and further select the literature collection of interest; (3) Literature exploration allows users to filter the literature through meta attributes in the attribute panel, explore the relations between literature at the topic level in the topic projection panel, and locate the relations at the publication level in the citation relation panel

2 Related work

The literature metadata contain multiple attributes (publication date, authors, affiliations, etc.). In addition, literature also contains textual data (abstract, keywords, etc.) and network data (citation, collaboration, etc.). According to the analysis target, we divide the literature visualization into three categories: text-based, network-based, and attribute-based (Fig. 1).

2.1 Text-based literature visualization

Textual data of literature contain titles, abstracts, keywords, and the main body. The textual information is the essential part of the literature and describes the main scientific contributions.

Considering the large literature dataset, users need to search for the literature of interest through technical terms. One of the most critical tasks of literature visualization is to help users search efficiently and visualize the searching results better. TileBar (Hearst 1995) splits the text into coherent parts of different topics based on the slicing techniques. Users can enter the terms of interest and their frequency requirements, and the system visualizes the searching results using a heatmap visualization to show the distribution of technical terms. The number of results could be enormous. Nowell et al. (1996) try to organize the results with a highly flexible visualization system. Users can select the metadata attribute of the literature as axes, such as authors and publication date, then the collection of search results is reorganized through a matrix view. CyBis (Costagliola and Cybis 2011) employs a 3D visualization technique. More specifically, users' searching results are displayed on a 2D plane according to their similarities. In addition, the publication date serves as the third dimension, which is convenient for users to explore the distribution of the literature in the searching results. The above studies only focus on a single searching result. Furthermore, Sparkler (Havre et al. 2001) enables the comparison for multiple searching results, which arranges the searching results along a circle, with each query result divided into a radial segment. The distances from the center encode the relevance between research studies and technical terms. Sparkler also uses color to encode the searching results of different queries.

In addition to filtering textual data according to user's queries, some research work also supports processing the textual data and further extracts research topics, which is essential for understanding research development. ThemeRiver (Havre et al. 2000) is a common visualization technique for studying the evolution of topics and arranges the temporal to the spatial dimension. ParallelTopic (Dou et al. 2011) also utilizes the Themeriver to show the topic evolution. In addition, it also focuses on the visualization of the topic modeling process for literature through the parallel coordinate system to show the distributions on different topics. According to the extracted topics, the literature is divided into different groups, and the

results can be visualized using several different approaches. PaperLens (Lee et al. 2005) utilizes the stacked bar charts to visualize the distributions of each year. Bubble sets (Collins et al. 2009) arrange the papers along the timeline and divides them into several groups using the colored bubbles. In addition to the visualization of static topics, Jiang et al. (2016) focus on the evolution along the timeline using a Sankey diagram and visualize the relationship among topics explicitly. Based on the extracted topics, Jigsaw (Görg et al. 2013) supports the flexible exploration of relationships between different meta-attributes (e.g., publication date, co-authorship, etc.) with the combination of textual analysis. The topics extracted from literature may also contain a hierarchical structure. HierarchicalTopic (Dou et al. 2013) creates topic hierarchies through topic modeling and hierarchical clustering algorithms, using a hierarchical topic river visualization to allow users to understand the evolution of topics and topic groups. With the development of deep learning techniques, more topic extraction methods are used for literature analysis. GaleX (Li et al. 2020) is designed for a larger-scale literature dataset and uses Doc2Vec (Le and Mikolov 2014) and t-SNE (Van der Maaten and Hinton 2008) techniques to analyze the evolution, intersection, and other aspects of topics in the literature at three progressively fine-grained levels: discipline, area, and institution. Furthermore, to improve the explainability of extracted topics, KeywordMap (Tu et al. 2021) develops a fine-tune Transformer-based neural networks, helping users identify important words and understand the keyword network structure within topics.

2.2 Network-based literature visualization

The literature dataset contains several networks, such as the collaboration network, the citation network and the co-citation network. Because the size of the literature dataset is often quite large, existing research studies focus on visualization techniques to avoid severe occlusions among nodes and edges.

Building a collaborative network among scholars can reach out a more comprehensive evaluation. According to the difference of entities in the network, we divide the network into three categories. The first category is the scholar-publication network. In the network, the nodes are either the authors or the publications. The links between the nodes are authoring relationship (Misue 2008). The second category is the collaboration network among scholars. The inner nodes encode scholars, and the links between nodes encode the collaborations between scholars. Huang et al. (2006) focus on the collaboration of a single author to avoid too much overlap between nodes. After selecting scholars interactively, the system builds an ego-centric collaborative network of scholars based on their published literature. The number of nodes and connected edges of the ego-centric network is significantly simplified, facilitating the design and implementation of visualization. Considering the varying density between different groups of scholars, Node-Trix (Henry et al. 2007) utilizes a hybrid visualization of node-link diagram and matrix to show the collaboration network. Shi et al. (2015) extend the central scholar of ego-centric network with a novel timeline-based glyph, and the links which represent the collaborations at the specific points of the timeline to visualize the evolution of the collaboration network among scholars. Furthermore, egoSlider (Wu et al. 2016) utilizes the glyph-based timeline visualization to summarize the evolution of a scholar's ego-centric network and also support side-by-side comparison.

The citation relationships between literature can also build networks. The analysis of citation networks can support a variety of tasks, including similarities of topics, influence propagation, research evolution. The entities of the citation network could be research papers (Garfield 2004; Matejka et al. 2012; van Eck and Waltman 2014; Chou and Yang 2011; Bergström and Whitehead Jr 2006; Dörk et al. 2012), scholars (Wang et al. 2018; Shin et al. 2019), and even conferences (Henry et al. 2007; Yoon et al. 2020). For the citation network based on research papers, Garfield (Garfield 2004) utilizes the node-link diagram along the timeline to show the complete citation network. Citeology (Matejka et al. 2012) complements with a stacked bar chart to show the visual summary, and CitNetExplorer (van Eck and Waltman 2014) defines the various types of nodes and edges and supports users to filter both nodes and edges of the network based on the network relationships to reduce the number of nodes in the network. Similar to collaboration networks, research studies (Chou and Yang 2011; Bergström and Whitehead Jr 2006; Dörk et al. 2012) also focus on the ego-centric network of one publication. PaperVis (Chou and Yang 2011) and CircleView (Bergström and Whitehead Jr 2006) put the focused publication at the center and utilize the radial layout of the node-link diagram to show the citation relationships. Pivotpaths (Dörk et al. 2012) enable to switch the focused entities flexibly. Many studies (Henry et al. 2007; Yoon et al. 2020; Wang et al. 2018; Shin et al. 2019) aggregate the literature dataset at the entity level to simplify the nodes of the citation network. The aggregations of the citation network improve the efficiency of visual layout and adds different perspectives

of explorations in the literature dataset. Existing studies utilize scholars (Wang et al. 2018) or conferences (Henry et al. 2007; Yoon et al. 2020) as the entities for aggregating literature, but do not utilize topics as the aggregating entities.

2.3 Attribute-based literature visualization

The attribute-based literature visualizations do not emphasize textual and network data characteristics but treat them as general attributes of literature (e.g., author and date). Many studies (Stasko et al. 2013; Heimerl et al. 2016) focus on the citation relationships between literature, but they emphasize the visualizations of multiple attributes instead of the structure of the citation network. CiteRiver (Heimerl et al. 2016) combines the content of publications and citation relationship together to provide a comprehensive analysis. PivotSlice (Zhao et al. 2013) proposes a multi-focus and multi-scale tabular view that provides progressive, flexible, and dynamic queries for the literature at different granularities. The literature contains diverse entities, and the intersection relationships between the entities are quite complex. Many analysis tasks are based on the overlapping relationships between entities. RadialSets (Alsallakh et al. 2013) presents a novel visual technique for radial sets, which can analyze the set intersections of a large number of elements. RadialSets utilizes a frequency-based representation to quickly find and analyze different kinds of overlaps between sets and correlate these overlaps with other properties of entities. While extracting topics from text data is a common approach, some work is based on labeled topics. For example, ConfVisExplorer (Chen et al. 2021) presents a new visual metaphor for analyzing and comparing the characteristics as well as the topic distributions of multiple conferences. Similar to our work, SurVis (Beck et al. 2016) also supports the literature management. Specifically, it enables users to create several filters from different aspects of meta-data and browse the filtering results. However, LitVis further provides a thorough analysis from text, network, and attribute aspects instead of straightforward browsing.

3 Requirement analysis

This section introduces the motivating design requirements of LitVis. To derive these requirements, we firstly summarized six procedures of writing literature reviews (search paper, identify important papers, read papers, build relationships, decide topics, and derive classifications) based on the guidelines proposed by McNabb and Laramée (2019). In addition, we worked closely with several faculties and graduate students to analyze challenges during the process of literature analysis and identify the following design requirements:

- *R1. Allow users to determine the literature collection of interest.* The literature contains both metadata and textual data attributes. In many cases, users need to filter the literature collection of interest from the whole literature collection based on these attributes, especially identifying the important papers. For example, according to the metadata attribute, users can select the publications of one scholar or several scholars. According to the textual data attributes, users can select publications of some specific topics.
- *R2. Support users in understanding the topics of literature collection.* After determining the literature collection of interest, users are required to read papers and summarize the topic distributions of these papers, which can facilitate users to understand the literature collection from the overview. In particular, users often have their own understanding about the publications. Therefore, it is also necessary to support users to adjust the topic of publications interactively.
- *R3. Enable to analyze the relationships among different topics.* Multiple relationships exist between different topics, including the relative chronological relationships, the intersection relationships, and the mutual influence relationships. It is required for users to analyze these relationships among the extracted topics. Furthermore, showing the relationships between topics at various granularities can also provide users with a detailed explanation.
- *R4. Provide the comprehensive analysis for the literature dataset.* In addition to the topics extracted from the textual data of literature, a comprehensive analysis for literature should combine with other metadata attributes, including author, venue, etc. Therefore, the prototype system should provide the distribution of the literature collection in terms of multiple metadata attributes and enable users' flexible interactions for the analysis.

4 System design

Driven by the above requirements, we propose a visual analytics approach for managing and exploring the literature dataset. Our system consists of two parts, literature management and literature exploration. In addition, this section also introduces the interaction designs of the prototype system.

4.1 Literature management

The literature management system allows users to upload the literature dataset in various data formats. As one of the basic formats for literature citations, bib files are supported by the majority of literature databases. Users can easily access bibtex data in these databases. In addition, our system also supports data in the JSON and CSV formats, which only need to be pre-processed according to the specific requirements with several basic fields. The benefit of these two formats is that it is easier for users to understand and manipulate. For example, users can add literature or modify the specific attributes manually.

The literature management system provides an overview of the entire literature uploaded by users and allows users to determine the literature collection of interest flexibly. As shown in Fig. 2, the visual designs of the literature management system are based on the visual mapping approach of CiteVis (Stasko et al. 2013) and extend it with more functionalities. The management system utilizes the unit visualization (Park et al. 2018). Each circle represents a publication, grouped by its publication date and venue along row and column. Each row indicates a venue, and each column indicates a publication date. The most recent publications are arranged on the left side. To support users identify the important papers, the density of circles encodes the number of their citations, which is one of the most important indicators of paper's importance. Specifically, the publications of darker circles have a larger amount of citations. Within each group of a specific venue and date, publications are sorted in descending order based on the number of their citations. To provide the distributions of other attributes in the overview, we encode the other metadata attributes into the stroke of circles. Figure 2 shows the comparison results of publication records from three different countries (China, USA, and Germany). A paper could be collaborated by several scholars from different countries, which is encoded by the segmented strokes of circles.

With interactive functions for searching and tagging, LitVis provide users with several functionalities about finding the literature collection of interest. Firstly, users can add tags or comments on the publications based on their understanding. Furthermore, by filtering the tags, users can accurately select the literature collection to be analyzed. During the analysis process, users can also modify the tags and comments of the literature on demand, thus updating the collection of analyzed literature (**R1**).

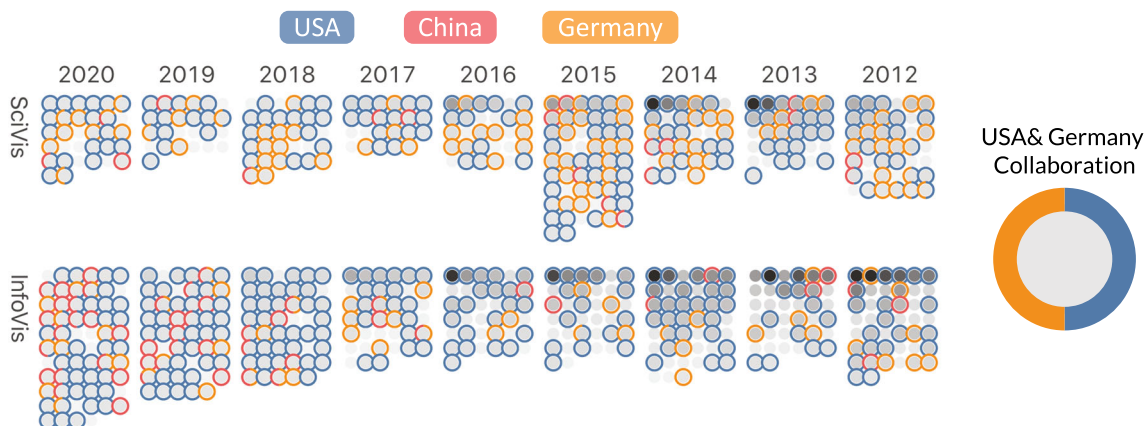


Fig. 2 The visualization design for the literature management. The left part shows the comparison results of publications from three different countries, USA, China, and Germany. The right part shows the glyph shows the publication collaborated by the authors from USA and Germany

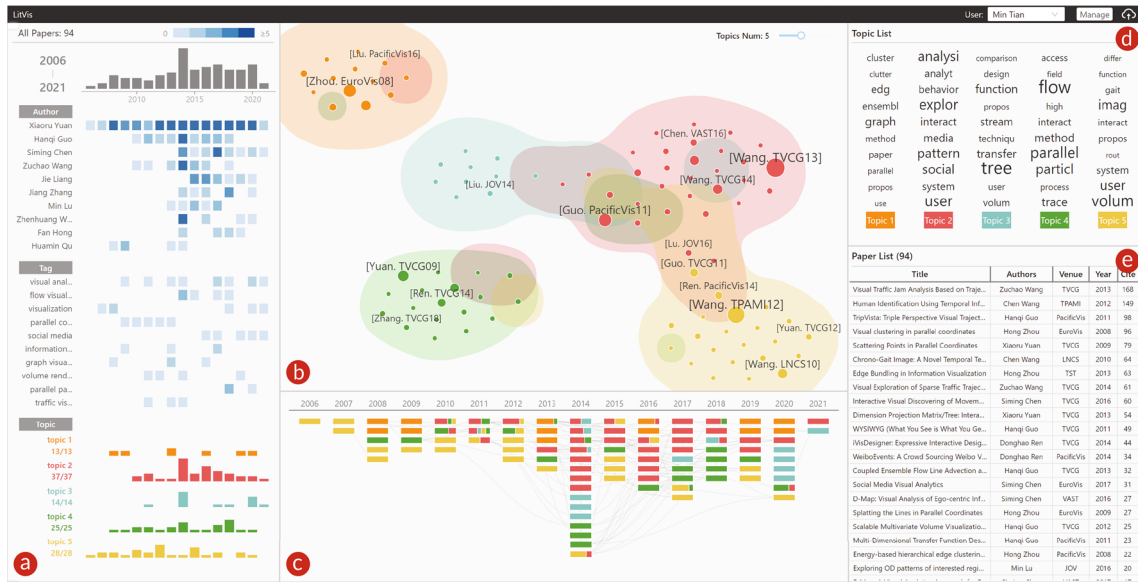


Fig. 3 The interface of LitVis prototype system comprises five interrelated components: **a** The attribute distribution panel, **b** The topic projection panel, **c** The citation relation panel, **d** The topic list panel, **e** The paper list panel

4.2 Literature exploration

We developed a visual analytics system to explore the selected literature collection of interest from attributes, text, and citation network aspects. As shown in Fig. 3, the interface for literature exploration consists of five panels. This section introduces the visual design of each component in detail.

4.2.1 Topic list panel

To facilitate users to understand the topic of the literature collection, we utilize the Latent Dirichlet Allocation model (LDA) (Blei et al. 2003) to extract topics from the literature. Compared with methods such as doc2vec (Le and Mikolov 2014) and BERT (Devlin et al. 2018), LDA does not need to be pre-trained with a large amount of dataset. In addition, another benefit of LDA lies in that the topic extraction results have a better differentiation and interpretability. Therefore, our method takes the title, abstract, keywords, user-annotated tags, and comments of the literature collection as input of LDA. More specifically, LDA computes a topic-word matrix, which provides the specific explanations for each topic. Each topic consists of multiple keywords, and LDA can compute the probabilities where each keyword belongs to different topics. By filtering the keywords with high probability, users can quickly understand the composition of the topic. LDA also computes a document-topic matrix that provides information about the associations between topics, and each paper can belong to more than one topic with different probabilities.

The topic list panel (Fig. 3d) provides detailed explanations of the topics (R2) by selecting the top 10 keywords with the highest probability and assigns different colors to these topics. In the topic list panel, the font size of words encodes its probability of belonging to the topic, to help users better understand the composition of the topic. Users can adjust the topic extraction results interactively. On the one hand, based on the topic extraction results, users can add tags or comments on the literature according to their understanding. On the other hand, we provide a slider for users to adjust the number of topics extracted by LDA.

4.2.2 Topic projection panel

The topic projection panel is designed to provide an overview of topics extracted from the literature collection. As explained in Sect. 4.2.1, the computation results of LDA contain the topic-document matrix, which provides the topic probability distributions of each literature. We extract the topic probability distribution as the representation of literature and then project it to 2D space using t-SNE (Van der Maaten and Hinton 2008). As shown in Fig. 3b, the projection result shows the overview of the literature collection.

More specifically, we map the literature into circles. The color of circles encodes the topic of literature. If a publication belongs to more than one topic, the color of the circle utilizes the color of topic with the highest probability. The size of circles encodes the number of publications' citations. The distances between circles encode the differences in the topic distributions of their corresponding publications.

Based on the projection results, we add more visual designs on the overview to facilitate users to understand the relationships among different topics. To help users understand the topic distributions and analyze intersections between different topics, we calculate the density contours for the literature of different topics. The density contour uses the same color as the topic but has a lower opacity to reduce their overlapping.

To help users understand the mutual influence relationships between different topics, LitVis aggregates the citation relationships between different publications at the topic level and add the node-link diagram based on the overview. As shown in Fig. 4b, when users select a topic as the focus to analyze the related topics, we add nodes at the geometric center of the selected topic and other related topics. The size of nodes encodes the number of papers within this topic. Note that each publication has one dominant topic but might belong to several topics with different probabilities. Therefore, each node is divided into two sectors, and the angle of sectors encodes the relative amount of different topics, showing the intersections of different topics. The links between nodes encode the influence relationships, which is calculated by aggregating the citation relationships between publications of different topics. The color of links encodes the direction of influence, which is consistent with the color of topics. For example, the links in red indicate the influences from the topics in red to other topics. The width of links encodes the size of influence, i.e., the number of citations between these two topics (R3). Note that when we calculate the size of influence, we exclude the intersection part of the literature and count only the citations of other parts.

4.2.3 Citation relation panel

The citation relation panel is designed to provide a more in-depth exploration for the influence relationship based on the literature perspective. As shown in Fig. 3c, we map publications into rectangles in the citation relation panel. The colors of rectangles encode the topics of the literature and the rectangles which indicate

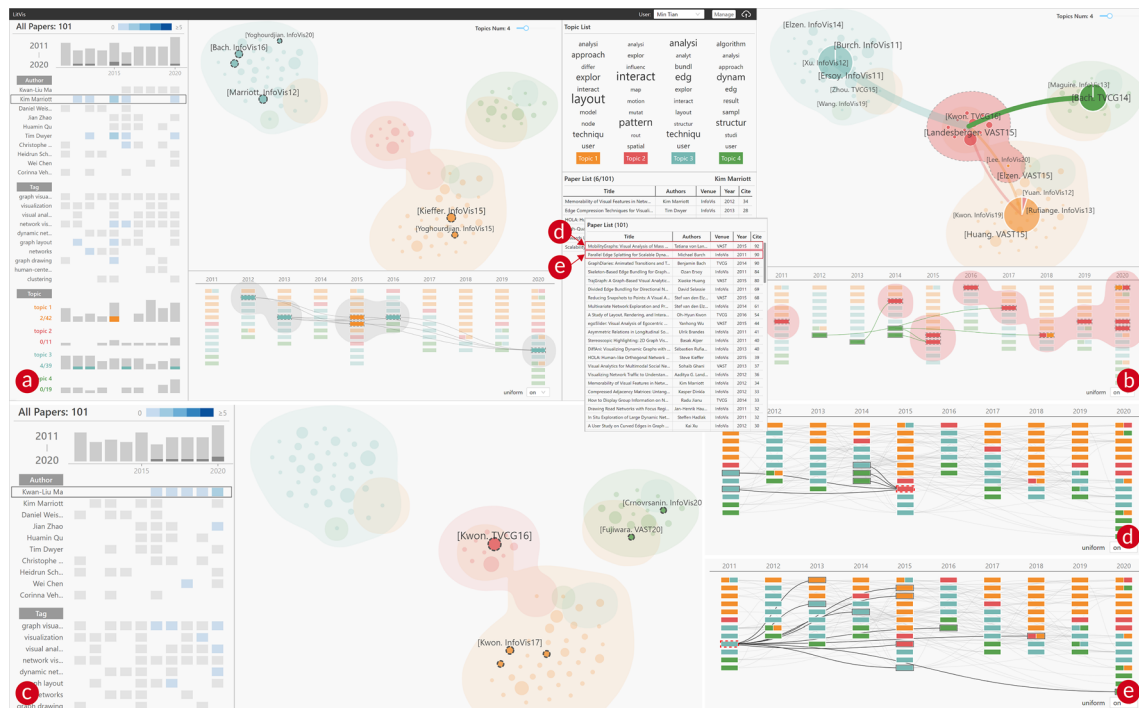


Fig. 4 Literature analysis for the publications related to graph visualizations of IEEE VIS from 2011 to 2020. **a, c** analyze the publications of an active researchers—Kim Marriott and Kwan-Liu Ma; **b** select the topic 2 as the focus and explore the mutual influence relationships between topic 2 and 4; **d, e** shows two different patterns of two highly cited publications

the publications with multiple topics are segmented into several *blocks* with corresponding colors. The probabilities of each publication belonging to different topics can be extracted from the topic-document matrix computed based on LDA. We provide two visual designs for the topic extraction results. The first design encodes the probabilities across different topics into the width of blocks as shown in Fig. 3c. Since the occupied space of each rectangle is small, the blocks which encode small probabilities are difficult to perceive by users. Therefore, the second design utilizes the uniform width for different blocks as shown in Fig. 4d. In addition, this visual design encodes the relative probabilities of different topics into the orders of different blocks.

In the citation relation panel, we layout the literature collection along the horizontal axis according to the publication date to show the development and decay of different topics. The rectangles are sorted along the horizontal axis according to the publication date. For the publication collection with the same date, we group the publications with the same topics to facilitate users to understand the temporal distributions of different topics. Each rectangle shares the same width and height. Based on the number of rectangles along both directions, the layout method calculates the height and width of each rectangle as well as its specific positions adaptively.

Comparing to the topic projection panel, the citation relation panel shows the references between literature at the publication level to support more fine-grained explorations of users. More specifically, it encodes the citation relationships between publications into the edges. When the citations between publications are very dense within the literature collection, the crossing of connecting edges will lead to severe cognitive burdens for the user's exploration task. To address this issue, the topic projection panel allows users to interactively highlight the rectangles and connected edges of interest in conjunction with specific exploration tasks as shown in Fig. 4b (R3).

4.2.4 Attributes distribution panel

The metadata attributes of literature have many dimensions, including authors, keywords, affiliations, journal, publication date. We provide the attributes distribution panel to support users to explore the relationships between different attributes (R4).

To support users in exploring patterns of different metadata attributes, the attribute distribution panel shows the overall attribute distributions of the selected literature collection along time. The evolution of the topics over time attracts a wide range of user interests. We also provide a bar chart of topics based on publication date to help users better compare their distributions, as shown at the bottom of Fig. 3a. Among all metadata attributes, users are usually more interested in the authors and topics of the publications, which can show the collaborations between different researchers. By processing the keywords and user annotations into the tag data, the attribute distribution panel presents the 10 most frequent scholars and tags and use the heatmap to visualize the temporal distributions of literature collection on these attributes, supporting users to visually compare different scholars as well as attributes.

The paper list panel (Fig. 3e) shows the entire literature dataset with metadata attributes by default, including title, first author, venue, publication date, and citation count. These information facilitate users to quickly locate literature of interest. When users select an author or tag of interest, the paper list panel will update the content accordingly.

4.3 Interaction design

In this section, we introduce the interaction designs of LitVis to meet the requirements of exploration and analysis.

4.3.1 Cross filtering

Many metadata attributes of literature, such as scholars, keywords, and annotations, often contain multiple values. For example, a publication is often collaborated by multiple researchers from several countries and consists of multiple keywords. When analyzing the literature dataset, it is essential to explore the relationships between different metadata attributes.

After determining the values of a specific attribute, such as scholar, tag, and topic, users can filter a subset from the whole literature collection. As shown in Fig. 4a, the attribute panel highlights the temporal distributions of the selected literature collection, allowing users to explore the distributions between

different attributes (**R4**). The citation relation panel highlights rectangles of literature with selected attributes or selected topics while generating an underlying area based on the density contour. The panel also shows all citation links inside the highlighted literature collection. Therefore, users can select the literature of interest for further exploration.

The topic projection panel will also update accordingly. When users select a specific attribute value of a feature, the topic projection panel highlights the corresponding literature. Moreover, when users select a topic as the focus of exploration, the topic projection panel will highlight the entire region of the topic, add nodes to indicate other topics, and show the influence relationships between the focused topic and other topics.

4.3.2 Influence exploration

With an overview of influence relationships between different topics, LitVis allows users to further explore the specific influence relationship at the literature level (**R3**). The main subjective of this functionality is to help users find the most valuable citation relationships based on their exploration tasks in a massive citation network, significantly reducing the cognitive burden of users. Users can select the link which indicates the influence relationship of interest at the topic projection panel, and the panel will highlight the links between the central topic and the topic of interest, as shown in Fig. 4b. At the same time, the citation relation panel below highlights the links between the publications of the focused topic and the publications of other topics to reduce the edge crossings. The color of the links is the same as the color of the source topic of the influence relationship. This design helps users to locate detailed information of the influence between different publications quickly.

5 Evaluation

To demonstrate the usability and effectiveness of LitVis, we invited 10 participants to use the prototype system, who are all graduate students and need to manage and analyze literature. The subjects are from different majors, including computer science, industrial design engineering, data science, etc. We present a usage scenario to demonstrate the effectiveness and usability of LitVis, which is based on the essential task of literature visualization, analyzing publications of a specific research topic.

We first introduced the functions of LitVis and the dataset used, and then encouraged the participants to explore freely according to their interests. At last, participants were asked to fill out a questionnaire about their experience of using LitVis and we interviewed each participant and collected their findings and comments.

5.1 Literature data collection

We asked the participants to analyze the literature collection related to graph visualizations, an important research field in information visualization. In recent years, researchers have proposed a series of general visualization techniques about graph visualizations, such as graph layout, and applied graph visualizations to various application scenarios, such as computer networks, social media. Therefore, analyzing and summarizing the development of graph visualizations is a practical task. To support users collect the papers related to graph visualizations, we provide the whole proceedings of IEEE VIS from 1990 to 2020. By searching for the related words, such as network, graph, participants can initially determine a collection of literature. Then they expand the literature collection according to the citation relationships between publications. After going through these publications, the participants can add tags to them manually for further filtering according to their understanding. Finally, participants can determine around 100 publications for further explorations.

5.2 User findings

We encouraged participants to share what they found in their free explorations and summarized these findings from three aspects: topics, scholars, and influence relationships.

5.2.1 Topics overview

Based on the literature collection collected by users, LitVis extracts four different topics, as shown in the Topic List Panel of Fig. 4a. These keyword distributions can help participants get an overview of the publications. More specifically, topic 1 and topic 3 are both related to “layout” and “technique” and the publications of these two topics occupy the majority of the whole collection (81/101). However, differences also exist between the publications of these two topics. Topic 1 emphasizes the novel “approaches” to optimize “model” and “node” layout, such as DiffAniRufiange and McGuffin (2013). Topic 3 focuses on the studies about the network “structure”, especially related to “edges” and “bundles”, such as skeleton-based edge bundling Ersoy et al. (2011). This finding inspires the participants to utilize visual elements of graphs as one dimension to analyze the existing techniques. According to this dimension, research studies can be classified into two categories. Research studies of the first category focus on the nodes of graph visualizations, such as improving the node layout algorithm or visual encoding approach. The second category emphasizes the edges between nodes in graph visualizations. One typical example is the edge bundling approach of graph visualizations (R2).

The keywords of topic 2 include “influence”, “map”, “motion”, “mutation”, “spatial”. These keywords facilitate participants to learn that the research studies of this topic emphasize the applications of graph visualization. Topic 4 has a striking keyword “dynamic”, which can be inferred that these studies are related to dynamic graph visualizations. Analyzing the evolution of large-scale graphs is one of the great challenges of graph visualization.

5.2.2 Active scholars

Users can analyze the relationship between different attributes and topics based on the attribute panel. As shown in Fig. 4c, the participants found that Kwan-Liu Ma is an active researcher in graph visualization and users wanted to analyze his research work in-depth with the topics projection panel. They found that Ma is mainly dedicated to research studies about graph layout, among which his publication on the immersive graph visualization Kwon et al. (2016) has a largest impact (R4).

Another finding is that Kim Marriott is one of a pioneer in the field of graph visualization research field, as shown in Fig. 4a. the participants found that Marriott has a lot of collaboration with Tim Dwyer (they both work at Monash University). After selecting the scholar, the participants found that his research work has many internal citations and inferred that there is a good continuity of their work on graph layout. These findings facilitate users read these papers, finally proved their hypothesis (R4).

5.2.3 Influence spread

The number of citations is an important aspect to evaluate the impact of a publication. Users sorted all the articles by citation count in the Paper List Panel and tried to analyze the highly cited papers. Parallel Edge Bundling has a lot of citations. As shown in Fig. 4e, it (Burch et al. 2011) has also been cited many times by the publications in the literature dataset, which informed the participants that this work made a significant contribution to graph visualizations as a general graph visualization technique. On the contrary, MobilityGraphs (von Landesberger et al. 2016) has a similar amount of citations with Parallel Edge Bundling, but it is not cited by the publications in this literature dataset. The participants read the paper’s abstract to figure out the reason and found that this work focuses on applying graph visualization methods to the study of mobility in traffic and transportation. The huge impact of MobilityGraphs shows that graph visualization has a huge impact on other research areas.

Users also investigated the influence relationships between graph visualization applications (topic 2) and others. They selected topic 2 as the focused topic, as shown in Fig. 4b, and found that topic 2 had very little output influence on other topics, but topic 3 and topic 4 had a very strong influence on it. Users clicked on the link which indicates the influence relationships between topic 2 and topic 4 to further explore the specific citation at the literature level and found that GraphDiaries (Bach et al. 2014) was cited twice. After reviewing the abstracts of these papers, users found that GraphDiaries and Segue (Law et al. 2018) are both related to dynamic networks. This finding informed the participants that GraphDiaries’ method on animation is helpful for the construction of the Segue system (R3). The users read the full text of these papers to validate their assumptions.

Table 1 Comparison between LitVis and existing literature visualizations

	Li et al., TVCG2020	Shin et al., VAST2019	Chen et al., JASIST2010	LitVis
Literature management	No	No	No	Yes
Aggregation level	Publication	Entity	Publication	Topic
Analysis aspects				
Metadata attributes	No	No	No	Yes
Textual data	Yes	No	No	Yes
Citation relationship	No	Yes	Yes	Yes

5.3 User feedback

All participants made more or fewer discoveries during the free exploration. They indicate LitVis is easy to learn ($\mu = 4.6$, $\sigma = 0.51$) and feel comfortable ($\mu=4.7$, $\sigma=0.48$). During the interview, several participants commented on the effectiveness: “*LitVis can give me some new insights.*” (P4), “*After I had made some discoveries, LitVis supported me in looking at more details to verify or disprove my hypothesis*” (P5) and usability: “*The system design is beautiful and easy to understand.*” (P3, P8, P10), “*The information displayed is very complete, and the temporal distribution of the attributes is helpful for analysis.*” (P9).

There are also some comments pointing out the parts that can be improved: “*Only support one author or tag for exploration. Multiple choice should be interesting.*” (P1), “*Although the citation links are reduced by interaction, there is still a lot of crossovers.*” (P7). Overall, the participants were satisfied with LitVis ($\mu = 4.2$, $\sigma = 0.63$) and indicated that they want to use it to manage and analyze papers of their field.

6 Discussion and future work

We compare LitVis with three existing research studies/commercial tools (Li et al. 2020; Shin et al. 2019; Chen et al. 2010) about literature visualizations from three different aspects, whether to enable the literature management, analysis aspects of literature visualizations, aggregation levels of citation relationships. We choose these research studies because they also focus on the citation relationship between publications. Table 1 shows the comparison results between LitVis and other related work. The results shows that LitVis is the only one which allows users to select the publication collection of interest through the literature management and analyze the literature collection from the metadata attributes, text data, and citation relationships aspects. In addition, LitVis also provides a novel perspective for aggregating the citation relationships at the topic level.

For the scalability, most visualization designs in LitVis can support the management and exploration of large-scale literature dataset except for the citation network view. Citation relations between literature are essential to help users understand their mutual influences between different research studies. These relations are also quite dense and complex. Visualizing the citation relationships directly will lead to severe occlusions. To visualize the citation relationships clearly, it is necessary to make simplifications, for example, filtering the ego-centric network based on a focused publication. Our method extracts topics and aggregates the citation relationships at the topic level, which can also be regarded as the simplifications for the citation relationships. The analysis task—understand the interplay among different topics—determines the simplification approaches for the citation relationship between different topics. However, when showing the detailed citation relationships between the specific publications within each topic, the current approach still has some occlusions between the citation links in the citation relation panel. In future, we plan to develop the visual design approach, especially the layout method, to reduce the occlusions.

LitVis allows users to upload a literature dataset and determine a collection of interest interactively, e.g., all publications of a researcher or related to a research topic. Furthermore, LitVis allows users to analyze the literature dataset from several perspectives to help the user understand the development of research studies. However, in some usage scenarios, the user may also be interested in two or even more papers collections. For example, users are interested in two researchers and want to make a comparison about their research studies, including the diversity of their research topic and evolution. It is an interesting research topic to support the user to compare multiple collections of literature.

Although our visualization methods are designed for visual analysis of the literature dataset. However, the problems and methods addressed in this study can also be generalized to other research topics. For example, the discussion of an event in social media may have several different topics that interact with each

other and the content of the discussion is constantly evolving, which is consistent with the scenario of our research work. Our research work can be subsequently extended to areas such as social media. In future, we plan to explore generic visualization methods of influence and evolution between topics to help users' understanding and exploration.

7 Conclusion

We propose LitVis, a visual analytic approach for literature management and exploration. LitVis allows users to select a literature collection of interest and make further exploration from three different aspects, attributes, text, and citation network. LitVis extracts the research topics and supports users to adjust the topic extraction results interactively. LitVis provides a novel perspective to show the interrelationship between various research topics and within the same topic at different granularities. We utilize the literature dataset related to visualization venue and present one use case to validate the effectiveness of LitVis. The result shows that LitVis is able to facilitate users to get valuable insights from the literature collection.

Acknowledgements The authors thank the anonymous reviewers for their valuable comments. This work is supported by National Numerical Windtunnel Project NNW2018-ZT6B12 and NSFC No. 61872013.

References

- Alsallakh B, Aigner W, Miksch S, Hauser H (2013) Radial sets: interactive visual analysis of large overlapping sets. *IEEE Trans Vis Comput Graph* 19(12):2496–2505
- Bach B, Pietriga E, Fekete J (2014) Graphdiaries: animated transitions and temporal navigation for dynamic networks. *IEEE Trans Vis Comput Graph* 20(5):740–754
- Beck F, Koch S, Weiskopf D (2016) Visual analysis and dissemination of scientific literature collections with survi. *IEEE Trans Vis Comput Graph* 22(1):180–189
- Bergström P, Whitehead Jr EJ (2006) Circleview: scalable visualization and navigation of citation networks. In: *Proceedings of the 2006 symposium on interactive visual information collections and activity IVICA*, pp 19–21
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Burch M, Vehlou C, Beck F, Diehl S, Weiskopf D (2011) Parallel edge splatting for scalable dynamic graph visualization. *IEEE Trans Vis Comput Graph* 17(12):2344–2353
- Chen C, Ibekwe-SanJuan F, Hou J (2010) The structure and dynamics of cocitation clusters: a multiple-perspective cocitation analysis. *J Am Soc Inform Sci Technol* 61(7):1386–1409
- Chen K, Wang Y, Yu M, Shen H, Yu X, Shan G (2021) Confvisexplorer: a literature-based visual analysis system for conference comparison. *J Vis* 24(2):381–395
- Chou J, Yang C (2011) Papervis: literature review made easy. *Comput Gr Forum* 30(3):721–730
- Collins C, Penn G, Carpendale S (2009) Bubble sets: revealing set relations with isocontours over existing visualizations. *IEEE Trans Vis Comput Graph* 15(6):1009–1016
- Costagliola G, Fuccella CV (2011) A novel interface for searching scientific documents. In: *International Conference Information Visualisation (IV)*, pp 276–281
- Devlin J, Chang M, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
- Dörk M, Riche NH, Ramos GA, Dumais ST (2012) Pivotpaths: strolling through faceted information spaces. *IEEE Trans Vis Comput Graph* 18(12):2709–2718
- Dou W, Wang X, Chang R, Ribarsky W (2011) Paralleltopics: a probabilistic approach to exploring document collections. In: *2011 IEEE conference on visual analytics science and technology (VAST)*, pp 231–240
- Dou W, Yu L, Wang X, Ma Z, Ribarsky W (2013) Hierarchical topics: visually exploring large text collections using topic hierarchies. *IEEE Trans Vis Comput Graph* 19(12):2002–2011
- Ersoy O, Hurter C, Paulovich FV, Cantareiro G, Telea AC (2011) Skeleton-based edge bundling for graph visualization. *IEEE Trans Vis Comput Gr* 17(12):2364–2373
- Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B, Vespignani A, Waltman L, Wang D, Barabási A-L (2018) Science of science. *Science* 359(6379):eaao0185
- Garfield E (2004) Historiographic mapping of knowledge domains literature. *J Inf Sci* 30(2):119–145
- Görg C, Liu Z, Kihm J, Choo J, Park H, Stasko JT (2013) Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Trans Vis Comput Graph* 19(10):1646–1663
- Havre S, Hetzler EG, Nowell LT (2000) Themeriver: visualizing theme changes over time. In: *IEEE symposium on information visualization (InfoVis)*, pp 115–123
- Havre S, Hetzler EG, Perrine KA, Jurrus E, Miller N (2001) Interactive visualization of multiple query results. In: *Proceedings of the IEEE symposium on information visualization (InfoVis)*, pp 105–112
- Hearst MA (1995) Tilebars: visualization of term distribution information in full text information access. In: *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI)*, pp 59–66
- Heimerl F, Han Q, Koch S, Ertl T (2016) Citerivers: visual analytics of citation patterns. *IEEE Trans Vis Comput Graph* 22(1):190–199

- Henry N, Fekete J, McGuffin MJ (2007) Nodetrix: a hybrid visualization of social networks. *IEEE Trans Vis Comput Graph* 13(6):1302–1309
- Henry N, Goodell H, Elmqvist N, Fekete J (2007) 20 years of four HCI conferences: a visual exploration. *Int J Hum Comput Interact* 23(3):239–285
- Huang T, Huang ML (2006) Analysis and visualization of co-authorship networks for understanding academic collaboration and knowledge domain of individual researchers. In: *Proceedings of third international conference on computer graphics, imaging and visualization (CGIV)*, pp 18–23
- Jiang X, Zhang J (2016) A text visualization method for cross-domain research topic mining. *J Vis* 19(3):561–576
- Klavans R, Boyack KW (2009) Toward a consensus map of science. *J Assoc Inf Sci Technol* 60(3):455–476
- Kwon O, Muelder C, Lee K, Ma K (2016) A study of layout, rendering, and interaction methods for immersive graph visualization. *IEEE Trans Vis Comput Graph* 22(7):1802–1815
- Law P, Wu Y, Basole RC (2018) Segue: overviewing evolution patterns of egocentric networks by interactive construction of spatial layouts. In: *2018 IEEE conference on visual analytics science and technology (VAST)*, pp 72–83
- Le QV, Mikolov T (2014) Distributed representations of sentences and documents. In: *Proceedings of international conference on machine learning (ICML)*, pp 1188–1196
- Lee B, Czerwinski M, Robertson GG, Bederson BB (2005) Understanding research trends in conferences using paperlens. In: *Extended abstracts of ACM Conference human factors in computing systems*, pp 1969–1972
- Li Z, Zhang C, Jia S, Zhang J (2020) Galex: exploring the evolution and intersection of disciplines. *IEEE Trans Vis Comput Graph* 26(1):1182–1192
- Liu S, Zhou MX, Pan S, Song Y, Qian W, Cai W, Lian X (2012) Tiara: interactive, topic-based visual text summarization and analysis. *ACM Trans Intell Syst Technol* 3(2):1–28
- Matejka J, Grossman T, Fitzmaurice GW (2012) Citeology: visualizing paper genealogy. In: *Proc ACM Conf Human Factors in Computing Systems (CHI)*, pp 181–190
- McNabb L, Laramée RS (2019) How to write a visualization survey paper: a starting point. In: *Tarini M, Galin E (Eds) Computer Graphics Forum*, pp 29–39
- Misue K (2008) Visual analysis tool for bipartite networks. In: *Proceedings of knowledge-based intelligent information and engineering systems (KES)*, pp 871–878
- Nowell LT, France RK, Hix D, Heath LS, Fox EA (1996) Visualizing search results: some alternatives to query-document similarity. In: *Proceedings of annual international ACM SIGIR conference on research and development in information retrieval*, pp 67–75
- Park D, Drucker SM, Fernandez R, Elmqvist N (2018) Atom: a grammar for unit visualizations. *IEEE Trans Vis Comput Graph* 24(12):3032–3043
- Rufiange S, McGuffin MJ (2013) Diffani: visualizing dynamic graphs with a hybrid of difference maps and animation. *IEEE Trans Vis Comput Graph* 19(12):2556–2565
- Shi L, Wang C, Wen Z, Qu H, Lin C, Liao Q (2015) 1.5d egocentric dynamic network visualization. *IEEE Trans Vis Comput Gr* 21(5):624–637
- Shin M, Soen A, Readshaw BT, Blackburn SM, Whitelaw M, Xie L (2019) Influence flowers of academic entities. In: *2019 IEEE conference on visual analytics science and technology (VAST)*, pp 1–10
- Small H, Boyack KW, Klavans R (2014) Identifying emerging topics in science and technology. *Res Policy* 43(8):1450–1467
- Stasko J, Choo J, Han Y, Hu M, Pileggi H, Sadana R, Stolper CD (2013) Citevis: exploring conference paper citation data visually. *Posters of IEEE InfoVis*, 2
- Tu Y, Xu J, Shen H (2021) Keywordmap: attention-based visual exploration for keyword analysis. In: *14th IEEE pacific visualization symposium, PacificVis 2021, Tianjin, China, April 19–21, 2021*, pp 206–215. IEEE
- Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. *J Mach Learn Res* 9(11):2579
- van Eck NJ, Waltman L (2014) Citnetexplorer: a new software tool for analyzing and visualizing citation networks. *J Inf* 8(4):802–823
- von Landesberger T, Brodkorb F, Roskosch P, Andrienko NV, Andrienko GL, Kerren A (2016) Mobilitygraphs: visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE Trans Vis Comput Graph* 22(1):11–20
- Wang Y, Shi C, Li L, Tong H, Qu H (2018) Visualizing research impact through citation data. *ACM Trans Interact Intell Syst* 8(1):1–24
- Wu Y, Pitipornvivat N, Zhao J, Yang S, Huang G, Qu H (2016) Egoslider: visual analysis of egocentric network evolution. *IEEE Trans Vis Comput Graph* 22(1):260–269
- Yoon T, Han H, Ha H, Hong J, Lee K (2020) A conference paper exploring system based on citing motivation and topic. In: *Proc IEEE Pacific Visualization Symposium (PacificVis)*, pp 231–235
- Zhao J, Collins C, Chevalier F, Balakrishnan R (2013) Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Trans Vis Comput Graph* 19(12):2080–2089

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.