CNVis: A Web-Based Visual Analytics Tool for Exploring Conference Navigator Data

Samuel M. Bailey; University of Notre Dame; Notre Dame, IN
Justin A. Wei; University of North Texas; Denton, TX
Chaoli Wang; University of Notre Dame; Notre Dame, IN
Denis Parra; Pontificia Universidad Católica de Chile, Santiago, Chile
Peter Brusilovsky; University of Pittsburgh, Pittsburgh, PA

Abstract

In this paper, we present CNVis, a web-based visual analytics tool for exploring data from multiple related academic conferences, mainly consisting of the papers presented at the conferences and participants who bookmark these papers. Our goal is to investigate the bookmarking relationships within a single conference and interpret various conference relationships and trends via effective visualization, comparison, and recommendation. This is achieved through the design and development of three coordinated views (the bookmark, topic, and keyword views) for user interaction and exploration. We demonstrate the effectiveness of CNVis using real-world data from three related conferences over a period of five years, followed by an ad-hoc expert evaluation of the tool. Finally, we discuss the extension of this work and the generalizability of CNVis for other applications.

Introduction

One of the main activities performed by research C ers in any field is synthesizing the knowledge created over time, presented as papers, through different theories, experiments, user studies, and related activities. This activity prevents researchers from reinventing the wheel and allows them to keep "standing on the shoulders of giants" to advance their fields, for instance, by analyzing trends of the research topics studied over time [18, 21]. The traditional artifacts used to present and disseminate these activities are survey papers, where a good deal of manual work is needed to search, gather, review, structure, and synthesize a collection of papers, software, technical reports, and even other surveys. One inconvenience of this form of presenting a large body of knowledge is that the authors must choose a specific way to structure and present the information, and other researchers might prefer making sense of that information in a different way, based on their needs. Progress in information retrieval, machine learning, web development, and information visualization has made possible the creation of tools that simplify several steps in the process of creating a survey and exploring a research field. The data used for these tools typically come directly from the textual content of papers, so co-authoring, co-citation, word co-occurrence are used as input for these analyses [7, 18].

In this paper, we present CNVis, an interactive visualization tool which uses textual and co-bookmarking data to allow users to explore a field, by means of entities such as authors, papers, and research topics automatically inferred using machine learning techniques. In particular, we analyze data from papers of several conferences and bookmarking activity of attendees over time to unveil important entities (papers, people), relations based on co-bookmarking (paper-paper, paper-attendee, attendee-attendee) as well as topics of interest over time. We used the data from the Conference Navigator website for this purpose.

Developed by the PAWS (Personalized Adaptive Web Systems) lab at University of Pittsburgh, Conference Navigator (CN) is a personal conference scheduling tool with social linking and recommendation features [31, 5]. The goal of CN is to enhance participant experience at conferences. It allows participants to supplement the basic functionality of a conference schedule with social networking, comments, and tagging. With CN, one can follow other conference attendees' schedules, make connections with new and old colleagues, and comment on talks and events. The system also generates general as well as personalized recommendations of papers (based on bookmarking) so people do not feel like missing the most relevant talks when attending the event.

Specifically, the CN data include the conference program data (conference sessions, papers, authors, abstracts, etc.) and the bookmark data made by participants. We clarify the terms "participant" and "user" which we use throughout the rest of the paper. A **participant** is a person who creates an account at the CN website and bookmarks papers that may interest him/her. The bookmark data serve as the key basis for our visual analytics tool. A **user** is a person who uses our visualization tool to examine the relationships between papers and participants. Based on the recommendations given, he/she may further bookmark papers before attending a conference, thus contributing to the bookmark data. In this sense, a user could also be a participant, or vice versa. In addition, users could be conference organizers who use our tool to identify the trend of a conference over years or investigate the similarities and differences among related conferences.

For this work, we collected the CN data for three related conference series (2011-2015): ECTEL (European Conference on Technology Enhanced Learning), Hypertext (ACM Conference on Hypertext and Social Media), and UMAP (International Conference on User Modelling, Adaptation and Personalization). Originally, the CN data were provided to us as files in the `csv` format. To make the data easier to manipulate and more compatible with our application, we process and convert these files to the `json` format. Specifically, we store the papers and participants for each year of each conference in a separate `json` file (e.g., all papers and participants for the ECTEL conference in 2011 are stored in one file). Each object in these files has the same structure, and is composed of three parts: the paper or participant's ID number, a data object holding the name of the paper or partici-

pant and a URL linking to more information, and a list of other participants or papers which the paper or participant is linked to. For example, a participant would have a list containing all of the papers he/she has bookmarked. Additionally, we create two more `json` files: one that assigns papers to the topic areas that they belong to, and another that stores keyword rankings for each topic, conference, and year.

Related Work

Previous works are related to our research in terms of 1) automatically obtaining the themes related to a scientific event or document collection, and 2) visualizing trends of a scientific event or domain with respect to their research topics. In this section, we review related work in these two areas.

Topic Detection in Scientific Documents

A topic can be defined as the subject matter of a document, talk, or discourse [24]. Different data sources and techniques have been used in order to automatically infer themes or topics from a collection of scientific documents. In terms of data sources, word co-occurrence (co-word analysis) coming from title, abstract, or keywords have been used to perform topic analysis using different techniques, such as the work of Isenberg et al. [21] which finds topics in the area of information visualization. Moreover, co-authorship and citation information has been used to model topics [11], study the relationships between researchers, and identify research communities [25, 27]. In terms of techniques, there are many options to infer topics from collections of documents. These options include, among others, clustering of terms after co-word analysis [21], self-organizing maps [34], and the popular latent Dirichlet allocation (LDA) [4], which has been used in several works to obtain topics from a collection of documents [18]. LDA can be seen as a technique evolving from latent semantic indexing (LSI) [10] and probabilistic latent semantic indexing (pLSI) [20]. LDA improves over LSI and pLSI by including Dirichlet priors in order to deal with restrictions, such as dealing with overfitting and being able to generate new documents (LDA is a generative model). Medlar et al. [26] introduced PULP, a system for exploratory search of scientific literature. They employed a temporal nonparametric topic model.

In our work, we use word co-occurrence from titles and abstracts of papers as the input for probabilistic topic modeling over time using dynamic topic models (DTM) [3], an extension of the popular LDA which accounts for the evolution of topics over time.

Topic Visualization

Topic visualization has emerged as an important research direction in visual analytics. Many works leveraged the river metaphor called ThemeRiver introduced by Havre et al. [19] to convey evolving topics over time. Cao et al. [8] developed FacetAtlas, a multifaceted visualization for entity-relational text documents. Their design visualizes both global and local relations of complex text document collections: global relations are displayed through a density map and local relations are conveyed through compound nodes and bundled edges. Dörk et al. [12] designed Visual Backchannel for following and exploring online conversations about large-scale events. Their design extracts keyword-based topics from tweets and provides an evolving, interactive, and multifaceted visual overview of large-scale ongoing conversations. Cui et al. [9] developed TextFlow, which leverages Sankey diagrams to visually convey topic merging and splitting relationships over time. Their approach extracts three-level features (topic evolution trend, critical event, and keyword correlation) and designs a coherent visualization to convey complex relationships between them. Dou et al. presented ParallelTopics [13] and HierarchicalTopics [14]. ParallelTopics uses ThemeRiver to display topic evolution over time and employs parallel coordinate plots to convey the probabilistic distribution of a document on different topics. HierarchicalTopics organizes the learned topics hierarchically and represents a large number of topics using Topic Rose Tree, showing the topic content as well as temporal evolution of topics in a hierarchical fashion. Alexander et al. [1] designed Serendip which includes a reorderable matrix, encoding of tagged text, and world ranking visualization to support multi-level serendipitous discovery in text corpora at multiple levels (the corpus, passage, and word levels). Oelke et al. [29] determined topics that discriminate a subset of collections from the remaining ones by applying probabilistic topic modeling and developed DiTop-View to visually compare content in multiple corpora. Gad et al. [16] designed ThemeDelta to investigate how trend keywords converge into topics and diverge into different topics. ThemeDelta also helps users identify temporal trends, clustering, and significant shifts in topics. Wang et al. [37] presented TopicPanorama, a visual analytics system for analyzing a full picture of relevant topics discussed in multiple sources (such as news, blogs, or micro-blogs). The visualization leverages a density-based graph layout along with a level-of-detail technique to balance readability and stability.

In this work, we perform dynamic topic modeling to identify topics from keywords extracted from papers of a conference series. We present a topic view using streamgraphs so that users can compare topic trends between conferences over time.

Task Analysis

Working closely with two domain experts in exploratory recommendation interfaces who are also co-authors of this work, we identify the following high-level functions or tasks users want to have when investigating the CN data:

T1. Overview of the data. This task aims to answer the question "*What do each year's conference and bookmark data look like?*" The CN data span multiple years, topics, and conferences. The visualization should provide an overview of the data and allow users to quickly and easily group or filter papers and participants in order to gain meaningful insights.

T2. Identification of popular papers and active participants. This task aims to answer the questions "*What are the popular papers?*" and "*Who are the active participants?*" It is common that in an academic conference, some participants are more active and some papers are more popular than others. Being able to visually identify these active participants and popular papers would allow conference organizers to predict future popular paper topics. This in turn would enable them to tailor fit the topics being discussed in papers at future conferences to be more in line with participant interests, raising conference attendance rates. Besides conference organizers, users or conference attendees would also be able to benefit from such information, which provides them hints to perform further exploration. For example, they would get ideas of which participants to follow and what paper presenta-

tions to join while attending the conference, extending the period of time they would spend there.

T3. Detail exploration of participant and paper attributes. This task aims to answer the question "*Can we drill down to a participant or paper of interest and obtain all related information?*" It is often desirable to know more about a particular paper or participant than just what is shown in a brief overview (i.e., paper title or participant name). Selecting a specific paper or participant and viewing a more detailed breakdown of their attributes (e.g., paper authors and abstracts, participant affiliations and URLs) would allow users to further explore papers or participants that they find interesting.

T4. Comparison of multiple participants in terms of papers bookmarked, and vice versa. This task aims to answer the questions "*What are the common and different papers bookmarked by multiple participants?*" and "*What are the common and different participants who bookmarked multiple papers?*" While not every participant will join the exact same list of paper presentations while attending a conference, it is likely to have overlaps among those lists. Given multiple participants, being able to view the similar and exclusive papers bookmarked would provide valuable insight into what types of participants bookmark what types of papers. On the other hand, seeing what papers are bookmarked by multiple participants would grant insight into why certain papers are more popular than others, and allow users to make correlations between paper topics and their popularity.

T5. Paper and participant recommendation based on user input. This task aims to answer the question "*Can we recommend similar papers and participants to users based on their input, such as paper keywords, topics, or the names of participants?*" Based on what papers a participant has already viewed or bookmarked, being able to suggest additional similar papers that might interest the same participant or other users can be very helpful. Papers could be suggested based on various metrics (topic, keyword, etc.), and would increase both participant attendance times and paper attendance rates.

T6. Comparison of the same conference data over years. This task aims to answer the questions "*Can we identify the trend of the same conference data over consecutive years?*", "*How do paper topics evolve over the years?*", and "*How does participants' involvement change over the years?*" Comparing data from a single conference over the course of multiple years would allow conference organizers to gain multiple insights into paper popularity and trends. For example, organizers would be able to track the popularity of paper topics and categories over time, allowing them to see popularity trends and even predict what topics would be popular in upcoming years. Attendance and involvement could also be tracked for participants who attended the conference multiple years, which would show changes in interest for specific participants. All of these insights would allow conference organizers to suggest future changes for a particular conference, boosting conference attendance.

T7. Comparison of multiple related conferences over years. This task aims to answer the questions "*Can we cross compare multiple related conferences over consecutive years?*" and "*What are the similarities and differences among them, in terms of both papers and participants?*" Extending the last task (**T6**), cross-comparing data from multiple related conferences over multiple years would allow conference organizers to see the similarities and differences between different conferences, in terms of both papers and participants. By identifying these similarities, organizers could co-locate or even merge conferences, both saving money for the organizations hosting the conferences and making it easier for participants to attend more paper presentations without having to travel as far or as often.

Design Requirements

Our visual analytics tool should meet the following design requirements in order to allow users to perform **T1** to **T7**.

R1. Visualize the connections among papers and participants. This requirement corresponds to **T1**, **T2**, and **T4**. In order to depict the connections between various papers and participants, the visualization should show connections between one or more participants and one or more papers (**T1**, **T4**) as well as help to identify particularly popular papers or active participants (**T2**).

R2. Provide separate rankings of papers and participants. This requirement corresponds to **T2** and **T3**. Providing a ranking system of popular papers and active participants can help users decide which participants to follow and/or which paper presentations to join while at a conference. This would allow them to view an overall list of popular topics or what active participants have bookmarked (**T2**), while also giving the option to view a more detailed description of each paper or participant (**T3**).

R3. Display detailed information associated with papers and participants. This requirement corresponds to **T2** and **T3**. Users may be intrigued by a particular paper's general overview, and wish to view either more detailed information about it or the paper itself. Linking this more detailed information to the ranked list of popular papers and active participants would allow easy access to both the general information of papers and participants (**T2**) and more in-depth descriptions of individual papers (**T3**).

R4. Enable participant-wise and paper-wise comparison and recommendation. This requirement corresponds to **T4** and **T5**. Once information concerning individual papers and participants can be displayed, the next logical step is to allow users to see the relationship between multiple papers and the participants that bookmarked them (**T4**). These comparisons could also form the foundation of a recommendation system in which users could input papers that they have already seen and receive recommendations in return on what other papers to view. Through topic modeling, papers can also be recommended based on a given keyword or a set of keywords (**T5**).

R5. Discover temporal and structural patterns of one or multiple conferences. This requirement corresponds to **T6** and **T7**. In order to satisfy the need for the comparison of conferences, both over time and with each other, our tool must handle all information connecting papers and participants to the conferences they were at. Specifically, two forms of visualization would need to be created: one to show multiple years of paper and participant data from a single conference (**T6**), and another to compare one or more years of data from multiple conferences (**T7**).

Preliminaries

In this section, we briefly introduce dynamic topic modeling and how to define the similarity or relevance between documents. In our scenario, a document is a paper.

Dynamic Topic Modeling

For each year of a conference, we extract keywords from the title and abstract of each paper. For all keywords we extract from a conference series, we generate *root words* (a root word is a basic word to which affixes are added to form new words) as new keywords. From these new keywords, we then perform *dynamic topic modeling* (DTM) [3], an extension of popular *latent Dirichlet allocation* (LDA) technique [4] to analyze the time evolution of topics in large document collections. The output of DTM is a list of topics for the conference series. Each topic, which is individually interpretable, is a probability distribution over keywords. Each paper is a mixture of topics.

According to [35], the probabilistic topic model specifies the following distribution over words within a document

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j), \quad (1)$$

where w_i is the i-th word, z_j is the j-th topic, T is the number of topics, $P(w_i|z_i = j)$ is the probability of word w_i under topic z_j, and $P(z_i = j)$ is the probability that topic z_j was sampled for word w_i.

Let Φ be the multinomial distributions over words for topics and Θ be the multinomial distributions over topics for documents. The parameters $\phi \in \Phi$ and $\theta \in \Theta$ indicate which words are important for which topic and which topics are important for a particular document, respectively. Blei et al. [4] introduced a Dirichlet prior on θ, as well as on ϕ, calling the resulting *generative model LDA*. As a conjugate prior for the multinomial, the Dirichlet distribution is a convenient choice as prior with hyperparameter α (for the case of θ) and parameter β_k (for the case of ϕ^k), simplifying the problem of statistical inference.

In practice, DTM works in a similar way as LDA, but it considers sequences of topics rather than static ones. LDA assumes that the order of documents does not matter. However, this assumption does not hold when analyzing collections that can span years, such as conference proceedings over consecutive years [2]. In such cases, we want a fixed number of topics changing over time, which is what DTM provides. Instead of a single distribution over words, a topic is now a sequence of distributions over words. Formally, the single Dirichlet hyperparameter β_k turns into $\beta_{t,k}$ in DTM, where the parameter evolves with Gaussian noise based on the previous state [3]

$$\beta_{t,k}|\beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I), \quad (2)$$

where t is a time slice and k is a topic. Similarly, the document-specific topic proportions θ are now drawn from a Dirichlet distribution with hyperparameter α_t, and the sequential structure between models is given by [3]

$$\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I). \quad (3)$$

By chaining together topics and topic proportion distributions, a sequential collection of topic models is obtained, with the k-th topic at slice t evolving from the kth topic at slice $t-1$.

Similarity or Relevance between Documents

With the set of topics derived from a conference series, we are able to compute the similarity or relevance between papers. Two documents (papers) are similar to the extent that the same topics appear in those documents. That is, the similarity between two documents d_1 and d_2 can be measured by the similarity between their corresponding topic distributions $\theta^{(d_1)}$ and $\theta^{(d_2)}$. In practice, we can use either the *symmetrized Kullback-Leibler divergence* or the *Jensen-Shannon divergence* as the similarity measure between probability distributions.

From information retrieval point of view, we can identify relevant documents by modeling information retrieval as a probabilistic query to the topic model. That is, the most relevant documents are the ones that maximize the conditional probability of the query, given the candidate document. Denoting $P(q|d_i)$ where q is a set of words contained in the query, we have

$$P(q|d_i) = \prod_{w_k \in q} P(w_k|d_i) = \prod_{w_k \in q} \sum_{j=1}^{T} P(w_k|z = j)P(z = j|d_i). \quad (4)$$

Note that this approach also emphasizes similarity through topics, with relevant documents having topic distributions that are likely to have generated the set of words associated with the query [35].

CNVis Tool

We develop CNVis as a web-based tool for exploring the CN data. The implementation uses D3.js for producing dynamic and interactive data visualizations in web browsers, along with utility functions provided by the jQuery JavaScript library.

As shown in Figure 1, our CNVis is made up of four components: the menu panel, bookmark view, topic view, and keyword view. The menu panel gives users an interface for interacting with participant, paper, and keyword data, and making comparisons between conference topics over the years. The bookmark view draws connections, for a given year, between the papers presented at a conference and the conference participants who bookmarked them. The topic view groups papers into topic areas based on their associated keywords. It displays topic popularity trends for a single conference or compares pairwise topic popularity between two different conferences over the years. Finally, the keyword view takes a list of keywords associated with a paper or topic and maps their popularity over the years. All these views are dynamically linked together via standard brushing and linking. Since the menu panel is self-explanatory, in the following, we only describe the three views in detail.

Bookmark View

As the main view of our CNVis tool, the bookmark view shows the connections between the papers and participants for a given conference at a given year. This view corresponds to the design requirements **R1**, **R2**, and **R3**. Through this view, users can either gain an overview of the connection among all papers and participants (by default) or a selected subset of either papers or participants and their corresponding connections (via filtering) for a single year of a given conference. The conference and year being displayed, as well as specific papers or participants, can be selected through the menu panel. In a circular layout, the bookmark view draws these connections as a bipartite graph (i.e., two distinct node sets: participant set and paper set) with edge bundling for visual clarity. We opt for the circular layout in order to display and explore a large number of papers and participants. Within the paper set, we further group the papers accord-

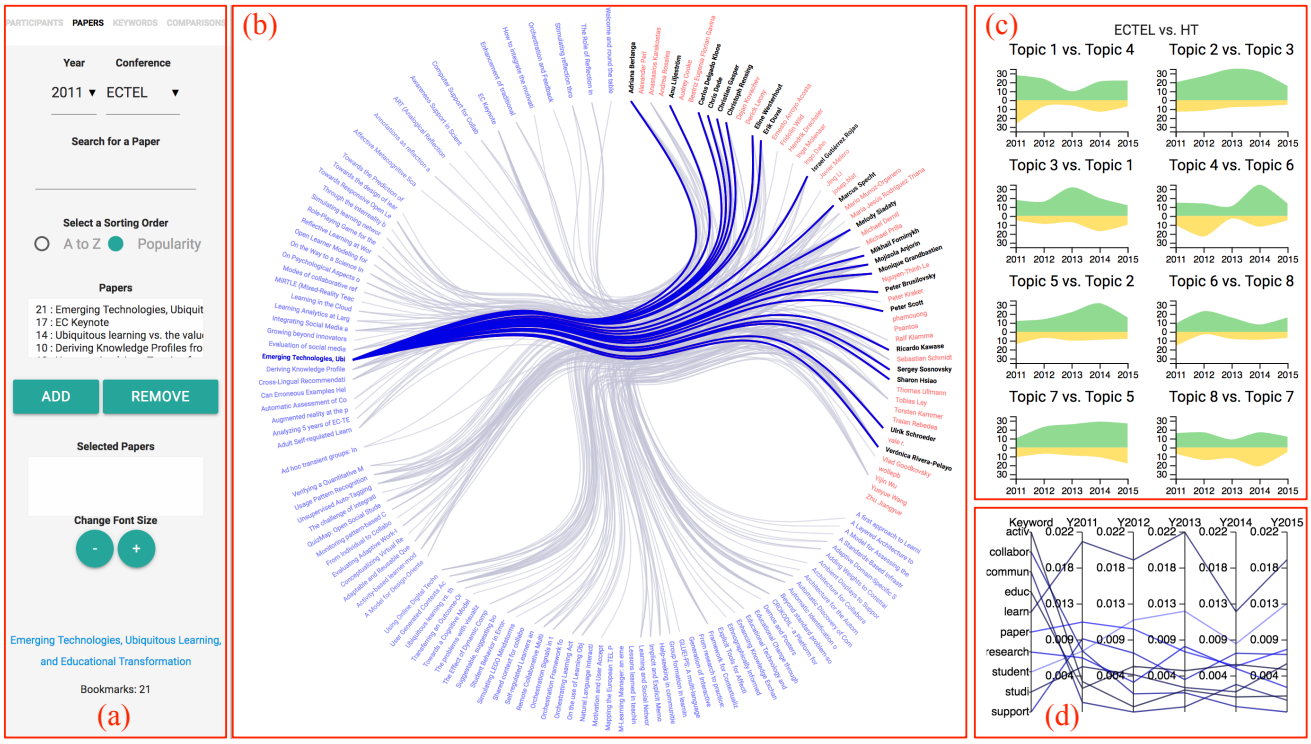


Figure 1. *The CNVis interface. (a) to (d) are the menu panel, bookmark view, topic view, and keyword view, respectively. The menu panel includes "Participants", "Papers", "Keywords", and "Comparisons" tabs. The bookmark view shows the connections between participants and papers using a bipartite graph with edge bundling. The topic view shows the pairwise comparison of conference topics and their trends using streamgraphs. The keyword view shows the top keywords associated with a paper or topic and their popularity over the years using parallel coordinates.*

ing to their similarities. The grouping is based on the clustering of the papers using *affinity propagation* [15]. We opt to use affinity propagation because unlike k-means and k-medoids clustering algorithms, affinity propagation simultaneously considers all data points as potential exemplars and automatically determines the number of clusters. As displayed in Figure 1 (b), we arrange the participant names alphabetically and within each group of papers, we arrange the paper names alphabetically as well. Paper groups are separated from each other with gaps along the circular layout.

Data Linking. As shown in Figure 1 (b), each link in the bookmark view represents a bookmarking connection between a participant and a paper. Paper titles are denoted in blue, while participant names are denoted in red. The links in the view also reflect this color coding: hovering over a participant's name highlights all papers bookmarked by the participant in red links, and hovering over a paper's title highlights all participants bookmarking the paper in blue links.

When selecting one or more papers or participants from the menu panel, all entities related to the selected entities are displayed in the *filtering mode* of the bookmark view. For example, if three participants are selected, the view updates to display those three participants and all papers that were bookmarked by one or more of those participants, and vice versa. In addition, when selecting specific papers or participants from the menu panel, entities can be sorted either alphabetically or by their popularity (i.e., the number of associated links), to allow users to either search more easily for a specific entity or simply select and view the most popular entities. If no papers or participants are selected in the menu panel, the bookmark view reverts to the *overview mode*.

Interaction. When hovering over a paper or participant around the periphery of the bookmark view, all links corresponding to the hovered entity (along with the entities being linked) are highlighted for easier user viewing. In addition, that paper or participant's detailed information is displayed at the bottom portion of the menu panel. This detailed information consists of the title of the paper or the name of the participant, and the number of bookmarks associated with the selected entity. The title or name of the selected entity is a URL, which sends users to a page displaying more information about the selected entity. Lastly, when clicking on a paper, the keyword view is also updated to display the top ten keywords associated with the selected paper and their popularity trends over the years.

Topic View

The topic view provides high-level pairwise comparisons between the distributions of the most similar topics throughout each conference. It corresponds to the design requirements **R4** and **R5**. The topic distributions are displayed through a series of small streamgraphs that share the same axes, so that users can clearly examine and compare the topic trends. Since our dynamic topic modeling outputs eight topics for each year of a conference, the streamgraphs are displayed in a 4×2 grid to facilitate comparison along both the year and quantity axes. Note that over the years, the set of keywords that define a topic for the same conference

may or may not vary, and the topics for different conferences are likely to be composed of different sets of keywords. Users may choose to display an overview of the topics of one conference or compare two different conferences at the topic level. They may also use the menu to indicate the subset of topics for display by adding and removing certain topics to the selection pool.

We determine topic pairs by first creating a distance matrix to record the distances among all topic pairs. For each pair of topics, we examine their keyword distributions and sum the probability differences of each keyword. If a keyword is unique, we add its probability value to the distance; otherwise, we add the absolute value of the difference between the corresponding probability values of the same keyword. We then use the *Kuhn-Munkres algorithm* to determine the set of pairs with the lowest sum of distances. Each topic pair is clearly displayed in the title of each streamgraph, where, in a comparison between the conferences ECTEL and HT in Figure 1 (c), the title "Topic 1 vs. Topic 4" refers to the first topic of the ECTEL conference and the fourth topic of the HT conference. We display the ECTEL topics in the positive (upper) section and the HT topics in the negative (lower) section of the streamgraph and color code the areas in green and yellow, respectively. Streamgraphs are sorted numerically by the topic IDs in the first selected conference (in this example, ECTEL).

Interaction. Users can hover over any part of each streamgraph to display next to the main title, the year, topic number, and quantity of papers corresponding to that respective area of the streamgraph. They can then click on an area of the streamgraph to select the subset of papers in that topic area in the bookmark view and the topic's keyword distribution in the keyword view. In this way, users are able to explore not only the recommended papers and their popularity related to each topic but also the detailed trends of each topic as explained by the corresponding keywords.

Keyword View

Given a list of keywords associated with a paper or topic for a given conference, the keyword view plots the popularity of those keywords to show their trends. This view corresponds to the design requirements **R3**, **R4**, and **R5**. It draws parallel coordinates to enable users to compare keyword popularity over the years, to both see past trends in keyword/topic popularity and predict what areas will be trending in the near future. Any number of root keywords can be searched for and selected from the "Keywords" tab of the menu panel. These keywords are displayed alphabetically along the "Keyword" axis of the parallel coordinates. The insights gained can help to tailor what papers should be featured at upcoming conferences in order to potentially increase participant attendance.

Interaction. The main interaction within the keyword view itself is axis brushing. By clicking and dragging on part or all of an axis on the parallel coordinates, only the results that run through the highlighted axis sections will be displayed in the keyword view and linked back to the bookmark view. This can be done for any number of axes, with a double click of a previously selected area removing the corresponding brushing. This feature allows users to hone in on specific keywords or years when looking at, for example, a list of the top ten keywords associated with a subset of papers.

Users can select one or multiple keywords in the "Keywords" tab of the menu panel. The tool will display in the bookmark view, only the related papers (following Equation 4) and the participants bookmarking one or more of them for the given conference at the given year. This interaction essentially recommends papers and participants based on selected keywords. The keyword view will also be updated to show the popularity trends of the selected keywords over the years.

Furthermore, the keyword view interacts with both the bookmark and topic views in multiple ways. If a single paper is selected in the bookmark view, that paper's top ten corresponding keywords (along with their popularity trends) are displayed in the keyword view. If multiple papers are selected from the "Papers" tab of the menu panel, the top keywords from each of the selected papers are used to calculate a keyword ranking via weighted averages, and the overall top ten keywords and their popularity trends are displayed. From these interactions, users are able to gain insights on similarities between any subset of papers from a given year of a conference.

In the topic view, if one of the areas is clicked on a streamgraph, the top ten keywords associated with the selected topic of the corresponding year and conference are displayed in the keyword view, as well as their popularity trends over the years. This allows users to indirectly compare individual keyword popularity trends between conferences. We note that for all interactions with the other views, only the top ten corresponding keywords are displayed in the keyword view. This is to keep the keyword view free from clutter and allow for easy and quick interpretation and understanding of the data being displayed.

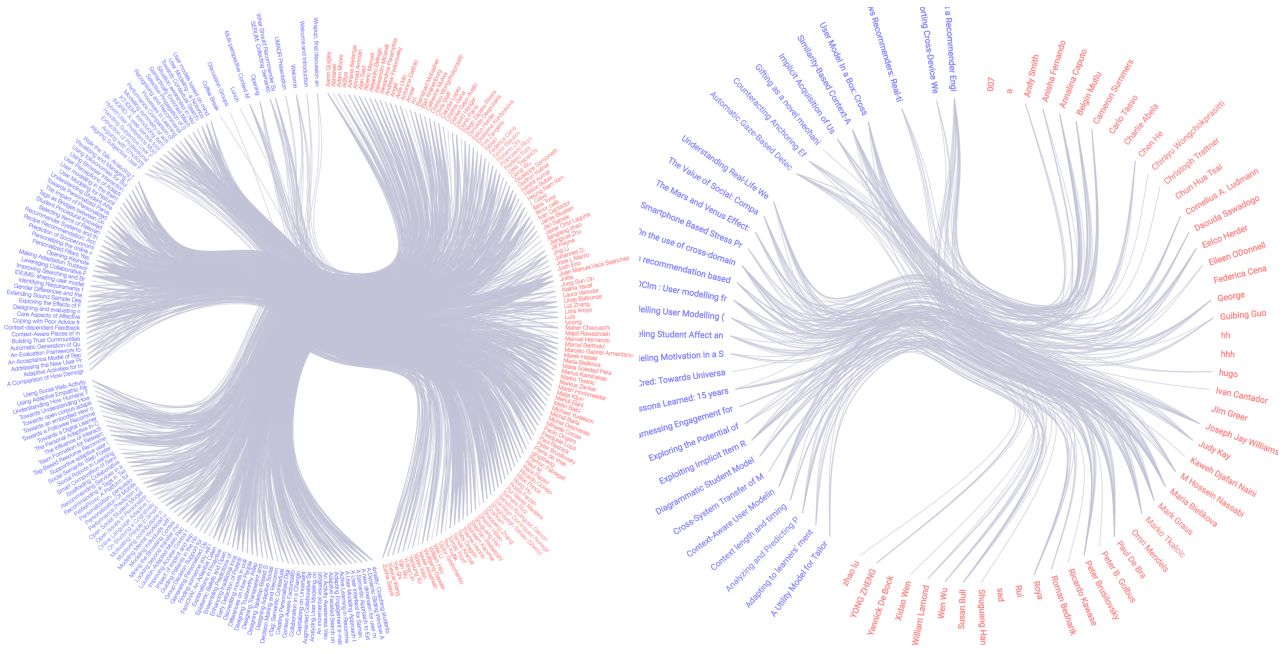
Results and Evaluation

Our CNVis tool is released online at: `http://sites.nd.edu/chaoli-wang/demos/`. To avoid any compatibility issues (known problems include the sorting by popularity), we recommend users to use the Google Chrome browser. In the following, we present three case studies and highlight the insights gleaned. The three studies jointly cover all seven tasks. Then, we report the evaluation we conducted with experts in recommender systems and human-computer interaction.

Case Studies

Case Study 1: Overview and Basic Selections. This case study addresses the needs of users to gain an overview of the data, select specific subsets of conference participants or papers, and examine detailed information about those specified participants or papers. Tasks **T1**, **T2**, and **T4** are covered here. We note that even though users in this case were only looking for more information on specific papers, the process for getting detailed information about participants is identical to that of papers. The only difference for participants is that clicking a participant in the bookmark view does not populate the keyword view (as participants do not have keywords associated with them).

In this study, users first wanted to see the overall volume of a conference for a particular year, and were able to do so by navigating to the "Papers" tab in the menu panel and selecting the desired conference and year. By then switching the selected conference and/or year, they were able to indirectly compare the overall size and popularity of one or more conferences. Figure 2 shows such a comparison. It is easy to see that UMAP 2011 has a significantly larger number of papers and participants than UMAP 2015. This



(a) (b)

Figure 2. *Comparing overviews of the UMAP conference. (a) shows a bookmark overview for 2011, while (b) shows the overview for 2015.*

stunning difference would promote users (such as conference organizers) to look further into why this decreased volume occurred over the years.

After gaining the overview, users then selected the specific papers or participants that they wished to explore, either sorting the list of papers or participants in alphabetical order or by popularity to facilitate their selections. In general, sorting in alphabetical order allows one to more easily search for a known paper or participant (although this can be completed by typing in the search bar in the menu panel as well), while sorting by popularity allows one to quickly identify the most popular papers or participants. By narrowing down the data being displayed in the bookmark view, users were able to view only the data that they deemed relevant and to examine the relationships between specific papers and participants. As shown in Figure 3, an example of this would be selecting the five most popular papers in a specific conference and year, displaying only those five papers (and their links to participants who bookmarked one or more of them) in the bookmark view, then examining the overlap between participant bookmarking groups. This would indirectly allow users to estimate the similarity between one or more papers based on the participants who bookmarked them. For instance, if the same participants bookmarked the same subset of papers, those papers may have similar topic areas or could be grouped together for the recommendation.

Case Study 2: View Interactions. This case study showcases the various interactions between the bookmark and keyword views of CNVis, and how those interactions lead to deeper insights gained from the data. Tasks **T3** and **T5** are covered here.

Users began this study by selecting a conference, year, and one or more keywords from the ”Keywords” tab of the menu panel. Upon this selection, they can then gain insight into the popularity trends of those keywords for the conference selected by viewing them over the years in the keyword view. Users have further potential to use these past trends to predict the popularity of trends in the future, and by linking these keywords back to the papers they are used in, users can gear future conferences towards paper topic areas that are predicted to draw in larger crowds.

Selecting these keywords also compiles a list of their top ten associated papers, displaying them along with the participants who bookmarked one or more of them in the bookmark view. Figure 4 shows such an example. This form of paper recommendation allows users to choose one or more keywords that they were interested in, and to have CNVis recommend what papers might be interesting based on their selection. This introduces users to new papers that they otherwise might not have found, to become more invested in the conference being viewed, and therefore explore it further. When users wanted to know more about a recommended paper, or about a participant who bookmarked one of those recommended papers, they could simply hover over the paper or participant around the periphery of the bookmark view to bring up detailed information. The included URL is useful for providing users with access to the papers being recommended, or more information about the participants who had already bookmarked them.

Case Study 3: Conference Comparisons. In the last case study, users wanted to compare the topic trends over the years, both within one conference and between two conferences, and to drill down from the overall topic trends to find recommended papers and keyword trends. These could answer why the conference topics were trending in that way. Tasks **T5**, **T6**, and **T7** are covered here.

Users first selected all eight topic areas of the ECTEL conference for both “Conf. 1” and “Conf. 2” from the “Comparisons” tab in the menu panel. From the resulting streamgraphs in the topic view, users were able to see that while some topic popular-

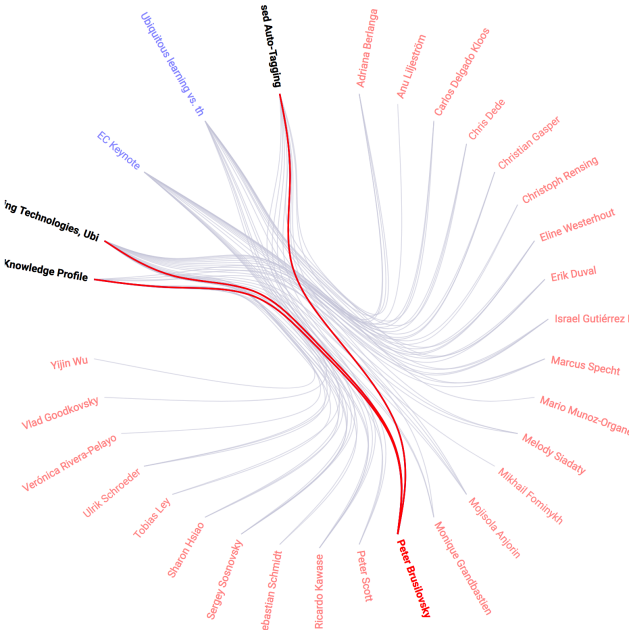


Figure 3. *The five most popular papers of ECTEL 2011 and the participants who bookmarked them. The papers bookmarked by Peter Brusilovsky are highlighted.*

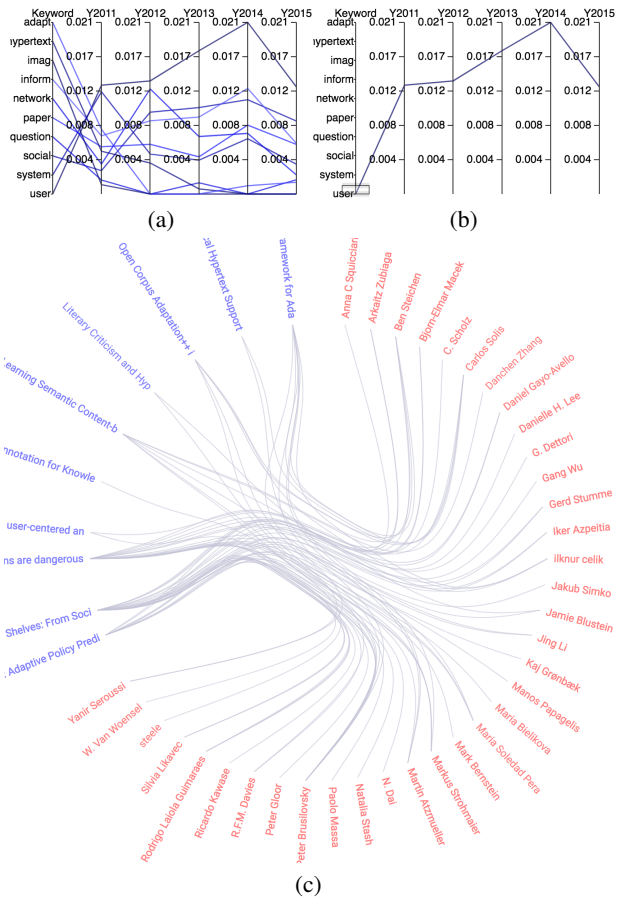


Figure 4. *(a) The top ten keywords associated with the HT 2011 conference paper entitled "Individual Behavior and Social Influence in Online Social Systems". (b) Axis brushing the keyword "user" in the keyword view. (c) The top ten papers associated with the selected keyword are shown in the bookmark view.*

ity trends were unrelated, others seemingly corresponded to each other. For example, in Figure 5, (a) and (b) show that Topic 1 saw a decrease in popularity in 2013, while in the same year Topic 3 saw an increase in popularity for a similar amount. This suggests that Topic 1's fall in popularity at that time could have, at least in part, been caused by Topic 3's rise in popularity. Moving on to comparing ECTEL's topic areas to those of UMAP, similar insights could be gained. Figure 5 (c) shows that the changes in the popularity of UMAP's Topic 5 could have been a contributing factor to the inverse popularity changes of ECTEL's Topic 1.

Wanting to explore the topics in more detail, users clicked on a section of one of the streamgraphs in the topic view to bring up all papers (and participants who bookmarked one or more of them) in the bookmark view that belong to the clicked topic in the corresponding year, as well as the top keywords associated with that topic in the keyword view. This gave users a complete overview of that topic for the year and conference selected. From this overview they were able to get a better understanding of what the topic was composed of, and when looking from one conference's topic composition to that of its most similar topic in the other conference, users were able to get an idea of why one topic's popularity in one conference could affect the popularity of the corresponding topic in the other conference. This comparison could be evaluated based on either the papers or the keywords, by simply examining the papers in the bookmark view or looking at the keywords associated with them in the keyword view.

Expert Evaluation

An expert in developing tools for analyzing data of conferences and research collections performed a heuristic evaluation of CNVis. Following a structure in the evaluation, the expert assessed the tool in the context of the seven tasks **T1**–**T7**. Afterward, the expert summarized his comments, both positive and critical, in three aspects.

T1–T4. When facing a task involving general exploration as well as detailed browsing, I usually expect to find the tool to able to comply with Schneiderman's information seeking mantra "*overview first, zoom and filter, then details-on-demand*" [32].

Pros. In terms of *overview*, the menu of CNVis allows me to select conferences, years, as well as interesting entities such as papers and participants. The bookmark view complements this menu selection in a constructive way, providing also the functionality for "*zoom and filter, then details-on-demand*". It smoothly allows visualizing at a high level, in a circular layout, the most popular papers, participants, as well as connections between these two types of entities. Colors seem to be chosen appropriately to discriminate between both types of entities, and the highlighting mechanism allows to hover over participants and papers to see the detailed information about these entities.

Cons. The main drawback I observe is the lack of authors as entities to explore in the menu panel and bookmark view. Being able to select well-known authors to consequently explore their papers is an interesting way to browse a collection of scientific

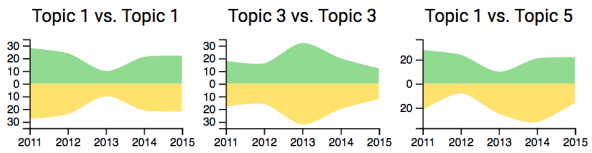


Figure 5. *Comparing popularity trends of either one or two conferences. (a) to (c) show the topic trends for ECTEL Topic 1, ECTEL Topic 3, and a comparison between ECTEL Topic 1 and UMAP Topic 5, respectively.*

documents, as shown by previous user studies on the original CN platform [31, 5]. In terms of visualization and interaction, the font size of the papers in the bookmark view seems too small. [We later added two buttons in the menu panel as shown in Figure 1, allowing users to change the font size in the bookmark view.] Although they can be seen in larger font at the bottom of the menu panel after hovering over them, exploring the detailed information of papers bookmarked by a single participant requires multiple interactions. A simple solution to implement could be: when clicking on a participant in the bookmark view, a list of all the papers bookmarked by this participant is displayed in the bottom portion of the menu panel.

T5. This task is about recommendation based on user input. Although CN has already a feature which recommends papers [6], and it has implemented interactive recommendation systems for specific conferences in the past [36, 30], none of them integrate the use of topic models to facilitate the exploration of participants and papers.

Pros. The brushing feature over the keywords in the keyword view is very useful to allow controllable exploration and recommendation, something that experts in a domain value with special emphasis [23]. The sorting mechanism (A to Z, popularity) on the menu panel which allows filtering the bookmark view entities, also acts as a step toward the recommendation. Finally, the linking between papers in the bookmark view and the keyword trends in the keyword view also provides an "inspection mechanism" to select interesting entities for further analysis.

Cons. The way that recommendation is implemented in the current version of CNVis does not learn a model of the user actions to eventually produce personalized recommendations. Recent works [17, 22] have shown that using reinforcement learning over user interactions on an interface could provide helpful feedback for personalized recommendations, and a future version of CNVis could be greatly enhanced by this addition.

T6 and T7. These tasks involve comparisons, either within conferences of the same series (in different years) or between different conferences, over time.

Pros. These are, in my opinion, the tasks better supported by CNVis in comparison to other tools surveyed in the related work. The streamgraphs and the use of topic models are excellent visual and interactive support tools to perform comparison tasks over a collection of documents using topics as proxies. The linking between streamgraphs and the keyword view also offers a useful mechanism to explore and compare within the same conference series as well as between conferences.

Cons. Sometimes there are keywords which are repeated between different topics, which hinders the comparison by a lack of clear discrimination between topics. This might be an effect of keywords which are too common over the whole corpus and hence, they show up as keywords with high probability in every topic. One way to alleviate this issue is the use of *relevance* of keywords within each topic, a concept explained by Sievert et al. [33] when implementing their tool named LDAvis.

Discussion

Privacy Concern

When users create an account in Conference Navigator, we disclose our data policy at `http://halley.exp.sis.pitt.edu/cn3/signup.php`, which states that:

Data Policy. *Conference Navigator is a research platform in which we study the ways to improve community-based and social recommendation systems. Data captured by the system includes your bookmarks, tags, social connections, logs, contributed external links, and metadata provided in your settings page. This data is essential to support social navigation and recommendation functionality of the system. The data will be kept confidential and will not be shared with any third party. No personal identifiers will be mentioned in any publications or dissemination of the research data. You can control information visible to other users of the system under your privacy settings in the profile page.*

Since CNVis is a tool intended to enhance "*social navigation and recommendation functionality*" of Conference Navigator, and we are not sharing users' data with any third party, we consider that we have addressed the potential privacy concerns of users.

Tool Validation

Ideally, the validation of a tool like CNVis should consider several types of evaluations. Following Munzner's nested model for visualization design and validation [28], one should consider validations at several levels: the domain problem (L1), data/operation abstraction design (L2), encoding/interaction technique (L3), and algorithm design (L4). In our case, we identified the tasks (L1), observed and interviewed target users (L1), justified the data/operation (L2) as well as encoding/interaction design (L3). We also empirically measured the performance of the algorithmic implementation (L4) and conducted an informal usability study (L3). Munzner's recommendation includes performing some validation of higher levels (L3, L2, L1) after implementation. Among them, we are still missing: a lab study to measure time/errors for operation (L3), a field study to document human usage of the deployed systems (L2) and finally, adoption rates (L1). Although these validations are very important, Munzner also states that "*Usually a single paper would only address a subset of these levels, not all of them at once*." In the future, we will conduct these additional validations (lab study, field study, adoption rates) to complete the evaluation of our tool.

Conclusions and Future Work

We have presented CNVis, a web-based visual analytics tool for exploring the CN data. Through interacting with a visual interface, we enable users to interpret various conference relationships and trends via comparison and recommendation using three coordinated views, namely, the bookmark, topic, and keyword views. The bookmark view allows users to examine, for a given conference of a year, the relationship between the papers presented at the conference and the participants who bookmarked them. The topic view allows for comparison of paper topic areas, either within a

single conference or between two different conferences, to reveal the overall conference trends. The keyword view enables the exploration of keyword popularity and their trends over the years for a given conference, either by selecting a specified subset of keywords or selecting one or more papers and viewing their associated keywords. We demonstrate the effectiveness of CNVis with selected case studies, followed by an ad-hoc expert evaluation of our tool.

The general framework of CNVis can be applied to other kinds of bookmark data, such as posts or images liked by social media users, products or services referred by customers, etc. We would like to explore this direction in the future. The key issue that needs to be addressed when extending and applying CNVis to other applications is the scalability (i.e., handling larger data while dealing with limited display). Given the limited screen space, the bookmark view typically could only display up to a few hundred entities. Beyond that, we may need to organize them into multiple levels of hierarchy in a sunburst view as an overview and use the bookmark view as the detailed view. Similar issues need to be addressed for the topic and keyword views, as appropriate.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation through grants IIS-1456763, IIS-1455886, and IIS-1560363. Denis Parra is supported by Conicyt Research Agency, Grant Fondecyt Nr. 11150783. We thank Lucas Barbosa, Matias Hurtado, Brendan Jones, Yike Ma, Charles Osborne, and Xin'an Zhou who helped with the project.

References

[1] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In Proceedings of IEEE Conference on Visual Analytics Science and Technology, pages 173—182, 2014.

[2] D. M. Blei. Probabilistic topic models. Communications of the ACM, 55(4):77—84, 2012.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In Proceedings of International Conference on Machine Learning, pages 113—120, 2006.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. The Journal of Machine Learning Research, 3:993—1022, 2003.

[5] P. Brusilovsky, J. S. Oh, C. López, D. Parra, and W. Jeng. Linking information and people in a social system for academic conferences. New Review of Hypermedia and Multimedia, 31 pages, 2016 (Accepted).

[6] P. Brusilovsky, D. Parra, S. Sahebi, and C. Wongchokprasitti. Collaborative information finding in smaller communities: The case of research talks. In Proceedings of International Conference on Collaborative Computing: Networking, Applications and Worksharing, pages, 1—10, 2010.

[7] M. Callon, A. Rip, and J. Law (Eds.). Mapping the dynamics of science and technology: Sociology of science in the real world. Springer. 1986.

[8] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. FacetAtlas: Multifaceted visualization for rich text corpora. IEEE Transactions on Visualization and Computer Graphics, 16(6):1172—1181, 2010.

[9] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. TextFlow: Towards better understanding of evolving topics in text. IEEE Transactions on Visualization and Computer Graphics, 17(12):2412—2421, 2011.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391—407, 1990.

[11] W. Ding and C. Chen. Dynamic topic detection and tracking: A comparison of HDP, Cword, and cocitation methods. Journal of the Association for Information Science and Technology, 65(10):2084—2097, 2014.

[12] M. Dörk, D. M. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. IEEE Transactions on Visualization and Computer Graphics, 16(6):1129—1138, 2010.

[13] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In Proceedings of IEEE Conference on Visual Analytics Science and Technology, pages 231—240, 2011.

[14] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. IEEE Transactions on Visualization and Computer Graphics, 19(12):2002—2011, 2013.

[15] B. J. Frey and D. Dueck. Clustering by passing messages between data points. Science, 315(5814):972—976, 2007.

[16] S. Gad, W. Javed, S. Ghani, N. Elmqvist, T. Ewing, K. N. Hampton, and N. Ramakrishnan. ThemeDelta: Dynamic segmentations over temporal topic models. IEEE Transactions on Visualization and Computer Graphics, 21(5):672—685, 2015.

[17] D. Glowacka, T. Ruotsalo, K. Konuyshkova, S. Kaski, and G. Jacucci. Directing exploratory search: Reinforcement learning from user interactions with keywords. In Proceedings of ACM Conference on Intelligent User Interfaces, pages 117—128, 2013.

[18] T. L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, 101(1):5228—5235, 2004.

[19] S. Havre, E. G. Hetzler, P. Whitney, and L. T. Nowell. ThemeRiver: Visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics, 18(1):9—20, 2002.

[20] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, pages 50—57, 1999.

[21] 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, 23(1):771—780, 2017.

[22] A. Kangasrääsiö, D. Glowacka, and S. Kaski. Improving controllability and predictability of interactive recommendation interfaces for exploratory search. In Proceedings of ACM Conference on Intelligent User Interfaces, pages 247—251, 2015.

[23] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. Inspectability and control in social recommenders. In Proceedings of ACM Conference on Recommender Systems, pages 43—50, 2012.

[24] O. Koltsova and S. Koltcov. Mapping the public agenda with topic modeling: The case of the Russian livejournal. Policy & Internet, 5(2):207—227, 2013.

[25] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. Information Processing & Management, 41(6):1462—1480, 2005.

[26] A. Medlar, K. Ilves, P. Wang, W. Buntine, and D. Glowacka. PULP: A system for exploratory search of scientific literature. In Proceedings of ACM SIGIR Conference on Research and Development in

Information Retrieval, pages 1133—1136, 2016.

[27] H. F. Moed. Citation analysis in research evaluation. Springer Science & Business Media, 9. 2006.

[28] T. Munzner. A nested model for visualization design and validation. IEEE Transactions on Visualization and Computer Graphics, 15(6):921—928, 2009.

[29] D. Oelke, H. Strobelt, C. Rohrdantz, I. Gurevych, and O. Deussen. Comparative exploration of document collections: A visual analytics approach. Computer Graphics Forum, 33(3):201—210, 2014.

[30] D. Parra, P. Brusilovsky, and C. Trattner. See what you want to see: Visual user-driven approach for hybrid recommendation. In Proceedings of ACM Conference on Intelligent User Interfaces, pages 235—240, 2014.

[31] D. Parra, W. Jeng, P. Brusilovsky, C. López, and S. Sahebi. Conference Navigator 3: An online social conference support system. International Conference on User Modelling, Adaptation and Personalization (Poster Paper), 4 pages, 2012.

[32] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In Proceedings of IEEE Symposium on Visual Languages, pages 336—343, 1996.

[33] C. Sievert and K. E. Shirley. LDAvis: A method for visualizing and interpreting topics. In Proceedings of ACL Workshop on Interactive Language Learning, Visualization, and Interfaces, pages 63—70, 2014.

[34] A. Skupin, J. R. Biberstine, and K. Börner. Visualizing the topical structure of the medical sciences: A self-organizing map approach. PLoS One, 8(3):e58779, 2013.

[35] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, (Eds.), Latent semantic analysis: A road to meaning. Laurence Erlbaum, 2007.

[36] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. Visualizing recommendations to support exploration, transparency and controllability. In Proceedings of ACM Conference on Intelligent User Interfaces, pages 351—362, 2013.

[37] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, and B. Guo. TopicPanorama: A full picture of relevant topics. IEEE Transactions on Visualization and Computer Graphics, 22(12):2508—2521, 2016.

Author Biography

Samuel M. Bailey *is a master student of computer science and engineering at University of Notre Dame. He received BS degrees in computer science and software engineering from Miami University in 2016.*

Justin A. Wei *is an undergraduate student of computer science at University North Texas. He conducted this work as an NSF DISC (Data Intensive Scientific Computing) REU student at University of Notre Dame during Summer 2017.*

Chaoli Wang *is an associate professor of computer science and engineering at University of Notre Dame. He received a Ph.D. degree in computer and information science from The Ohio State University in 2006. Dr. Wang's main research interests are scientific visualization and visual analytics.*

Denis Parra *is an assistant professor of computer science at Pontificia Universidad Católica de Chile. He received a Ph.D. degree in information science from University of Pittsburgh in 2013. Dr. Parra's research interests are recommender systems, intelligent user interfaces and information visualization.*

Peter Brusilovsky *is a professor of information sciences at University of Pittsburgh. He received a Ph.D. Degree in computer science from Moscow State University in 1987. Dr. Brusilovsky's research interests include adaptive web-based systems, personalized e-learning, adaptive hypermedia, social computing and social web, intelligent tutoring systems and shells, student and user modelling, and adaptive interfaces. He leads the PAWS lab which creates, develops, and maintains the Conference Navigator online system.*