Topic Modeling of Document Metadata for Visualizing Collaborations over Time

Francine Chen
FX Palo Alto Laboratory
3174 Porter Drive
Palo Alto, CA 94304 USA
chen@fxpal.com

Patrick Chiu
FX Palo Alto Laboratory
3174 Porter Drive
Palo Alto, CA 94304 USA
chiu@fxpal.com

Seongtaek Lim
School of Information
UC Berkeley
Berkeley, CA 94720 USA
stlim@berkeley.edu

ABSTRACT
We describe methods for analyzing and visualizing document metadata to provide insights about collaborations over time. We investigate the use of Latent Dirichlet Allocation (LDA) based topic modeling to compute areas of interest on which people collaborate. The topics are represented in a node-link force directed graph by persistent fixed nodes laid out with multidimensional scaling (MDS), and the people by transient movable nodes. The topics are also analyzed to detect bursts to highlight "hot" topics during a time interval. As the user manipulates a time interval slider, the people nodes and links are dynamically updated. We evaluate the results of LDA topic modeling for the visualization by comparing topic keywords against the submitted keywords from the InfoVis 2004 Contest, and we found that the additional terms provided by LDA-based keyword sets result in improved similarity between a topic keyword set and the documents in a corpus. We extended the InfoVis dataset from 8 to 20 years and collected publication metadata from our lab over a period of 21 years, and created interactive visualizations for exploring these larger datasets.

Author Keywords
Interactive visualization; small group collaboration; temporal dynamics; topic modeling.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Effective methods for data analysis and visualization of collaborations over time have many useful applications. They can be used for finding people with expertise in some topic area, people with broad interests across disciplines, and people who are recently active or who have had a large amount of experience over time. Other applications include helping management assess the balance of people working on various areas, and to see the impact of their policies on how their team members collaborate over time.

Temporal analysis of group collaborations within an organization can be a difficult task. One reason is that groups dynamically change over time in nonlinear ways [3]. For example, at a research laboratory, people with different skills and interests may collaborate with each other only for a few short-term projects. Thus, a means to analyze the dynamic features of collaboration would be useful.

Document metadata that contains co-author information provides a way to establish the relationships between people in collaborations. Along with who is involved, it is also important to know what they are collaborating on. Sometimes the metadata has information about publication venues (e.g. [17]), which can be used to define areas of interest. However, this is problematic when the dataset is focused on a single venue (e.g. [7]), or when scaling to a dataset with a large number of venues.

Another simple way to define areas of interest is to match the keywords from the document metadata to establish the relationships. The problem with this is that for a given keyword, the number documents (and authors) matching that keyword is relatively small. Hence a large number of keywords would need to be visualized (along with the authors), which is hard to do effectively.

This paper proposes the use of topic modeling to automatically compute areas of interest from document metadata for visualizing collaborations over time. Topic modeling provides more meaningful areas of interest than publication venues. A well-known topic modeling algorithm is Latent Dirichlet Allocation (LDA) [4]. As a dimension reduction technique, it enables better scaling to larger data sets than publication venues or keywords. Unlike simpler venue or keyword metadata, topic modeling provides a vector space of terms which has standard methods, such as cosine similarity, to compute relevance. Moreover, topic modeling enables filtering nodes by

IUI'16, March 07 - 10, 2016, Sonoma, CA, USA

ACM 978-1-4503-4137-0/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2856767.2856787

relevance, which has uses that include focusing on the most relevant nodes and reducing clutter.

Although the topic-model vector space is a reduced space from the high-dimensional vector space of terms, it needs to be further reduced for visualization. For this, the topic-model vector space provides a good way to lay out the topics in the visualization: the similarity scores between topics can be mapped to a 2D plane by multidimensional scaling (MDS).

Another important aspect for people who analyze data is to be able to see sudden surges in activity level. To visualize the changing popularity for the various topics over different time spans, we apply a burst detection algorithm [13] to the activity stream of each topic. The topics change colors to indicate different "heat" levels as different time intervals are selected by the user.

Our visualization interface employs a node-link graph with both fixed and moveable nodes, in conjunction with a time slider. Topics computed from the dataset are represented by persistent fixed nodes, which are placed in the 2D plane to indicate the relative similarity of the set of topics. People who work on different topics at different times are represented by transient moveable nodes, which are placed to indicate the similarity of their publications during a specified time interval to each other and to the fixed topic nodes. As the time slider is manipulated by the user to select a time period, the person nodes may appear, disappear, or move relative to each other and to the fixed topics with animated transitions.

To evaluate the effectiveness of using topic modeling in the visualization, we tested it on a public dataset (InfoVis 2004 Contest [7], extended to 20 years), and a dataset from our lab's publication database (21 years). We compared LDA-based topics to the topics others have automatically or manually defined for the public InfoVis dataset. For both datasets, we examined how well the topics cover the person nodes and the amount of clutter (number of nodes and edges) under different numbers of topics and relevance score thresholds, and the burst activity levels which can be affected by the diversity of topics.

RELATED WORK

In our previous work, Collaboration Map [18] demonstrated a visualization that uses a node-link force directed graph with fixed nodes representing publication venues and moveable nodes representing authors, along with a time interval slider. The use of publication venues is problematic when there are too few venues to be discriminative or too many venues to visualize and therefore does not scale.

In the present work, areas of interest are represented by topics; whereas in [18] the publication venues serve as surrogates for areas of interest. With our approach, the visualization model can be applied to document metadata focused on a single publication venue (e.g. [7]) or scaled to a large number of publication venues by controlling the number of topics. The topics are represented as fixed nodes that are laid out using MDS. Additionally, we perform burst detection on the topic streams to detect highly active topics.

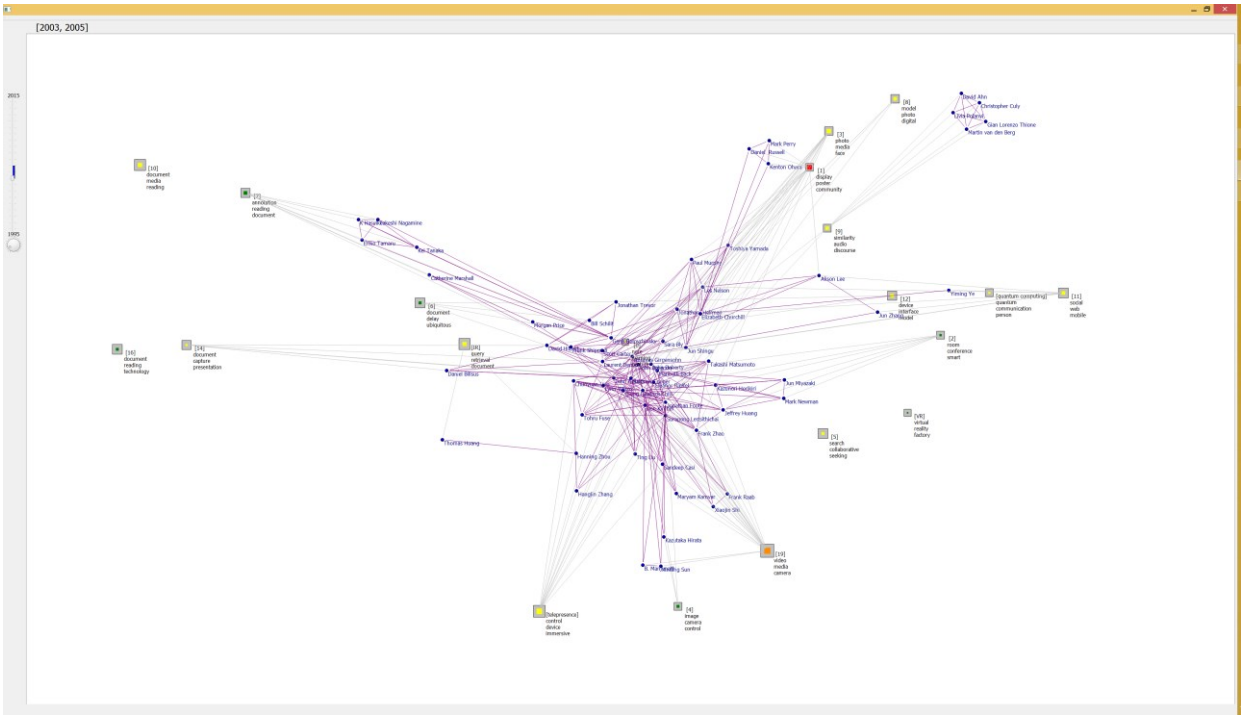
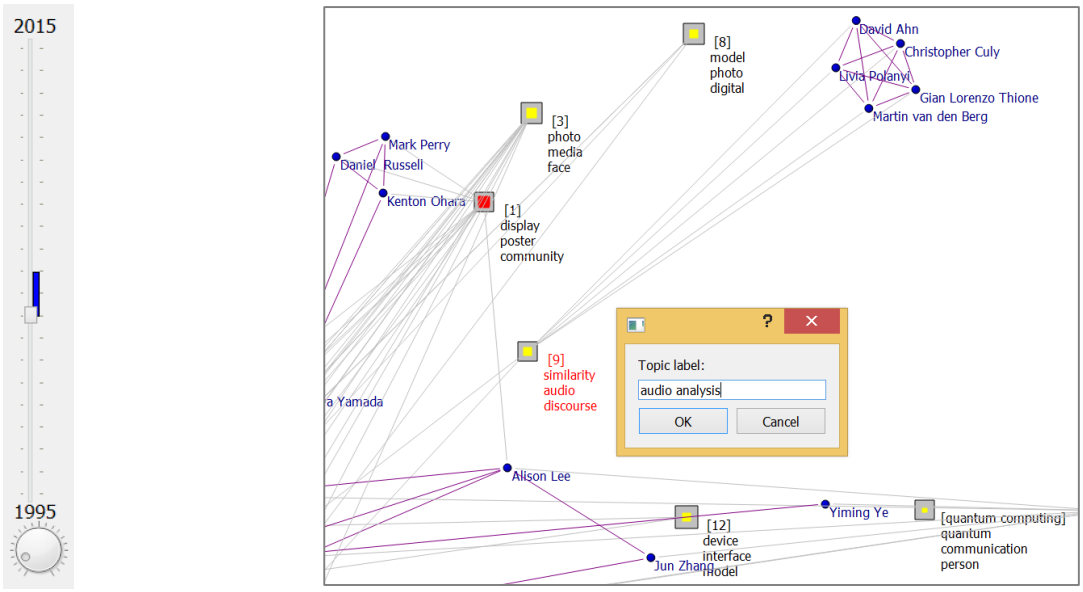


Figure 1. Screenshot of visualization showing topic-based collaborations. The dataset is from our research lab's publication database (Lab21), with 20 topics and relevance score threshold of 0.25. The visualization is rendered on a high-res 4K display.



(a) (b)

Figure 2. Close-ups: (a) timeline widget with a slider and a dial to set the interval size, (b) detail showing co-author relations, person-topic relations, topics and burst levels in the selected time period. A pop-up widget for adding a topic label is shown in (b).

Kraker et al. [14] visualize topic evolution using small multiples where at each time step the topics are computed and visualized; this results in the topic locations changing over time. In contrast, the topics are fixed in our visualization and serve as anchors for locating people. Furthermore, active topics are detected and shown in our visualization using different heat levels.

Most of the traditional visualizations for paper publications show a view with nodes representing the papers; examples are the InfoVis 2004 Contest entries [7] and the more recent PivotSlice [25]. In contrast, our visualization focuses on visualizing the collaborations and does not show the papers as nodes.

From the InfoVis 2004 Contest, the entry by Ke et al. [11], which was given a first-place award, has a node-link view of the authors in addition to a view of the papers. The co-author view shows only the authors (without topics or papers) and their relationships in a static view. In contrast, our visualization shows the authors' expertise by displaying links to the topic nodes, and also their co-author relationships for selected time periods in a dynamic interactive view with animations.

Regarding the data analysis aspect of collaboration networks and bibliographic metadata, existing work tends to focus on the theoretical properties of the network graphs (e.g. [16]), which are interesting for studying large scale systems and can help in the design of networking infrastructure, but are less applicable to networks in smaller organizations or teams.

The groups participating in the InfoVis 2004 Contest proposed a number of different methods for identifying keywords characterizing topics, which we refer to as *topic keywords*[1], to use in visualization of the InfoVis dataset. Both Ke et al. [11] and Tyman et al. [23] manually defined topic keywords. Teoh and Ma [22] identified keywords that most frequently occur in the documents in the dataset. Ke et al. [11] took into account temporal information and identified bursts of keyword usage.

Lee et al. [15] used Microsoft proprietary clustering to identify topics for the InfoVis 2004 Contest. Ahmed et al. [1] used a self-organizing-map to cluster documents. Identifying topics using these clustering methods results in a hard clustering of the documents, i.e. assigning each document to one cluster, where documents with similar term distributions are grouped.

We chose instead to use a soft two-sided clustering, Latent Dirichlet Allocation (LDA) [4], a popular method for identifying topics in a document collection. In the LDA model, each document is a mixture of topics, rather than being associated with a single topic, and each word in a document comes from one of the document's topics. Since documents often discuss more than one topic, the LDA model can better reflect the topics in each of the documents in a collection. We examine this in our experiments

[1] Note that a *topic keyword,* which summarizes a topic, should not be confused with keywords in the document metadata.

comparing the InfoVis 2004 topic keywords and LDA-based topic keywords.

The Termite system [5] is a visual analysis tool that provides a table visualization of terms and topics computed using LDA to help users to assess topic model quality. This is useful when a large number of topics is required, since the topic terms are often less coherent and "junk topics" become a problem.

DATA ANALSYSIS AND VISUALIZATION

An overview of the system is shown in Figure 3. We explain these parts in the following sections.

Data Preparation

To prepare a dataset of documents for use in the visualization, the document metadata is preprocessed and analyzed for topics. The results are represented with data structures that can be quickly initialized and run in real time. Typically, the document metadata contains the authors, time information (e.g. year), title, and keywords. Additionally, we require an abstract, because some amount of text is needed for the topic modeling.

Data Preprocessing

The metadata (abstract, keywords, title) is lower-cased and then tokenized into words with punctuation stripped at word boundaries. Stopwords are removed and then a singularizer, which converts plural word forms to singular, is used to normalize the remaining words. These "cleaned" words form the corpus vocabulary and create a vector space of terms V. Each document in the collection, D, is represented as a vector of term counts in V.

Topic Modeling

From the vector representation of documents, topics are identified for use as fixed interest areas in the visualization. As discussed in the Related Work section, there are many approaches to identifying topical keywords, but we use an LDA model, which can better represent the topics. The LDA model is trained on the preprocessed document data.

Given the number of topics desired, k, and the set of documents D represented as term vectors in V, Latent Dirichlet Allocation computes a set of k topics, where each topic is represented by a set of terms and their associated probabilities. We used the Gensim Python software library [8] for this computation. In the visualization, k is predetermined (or alternatively k can be chosen by the user).

For a given dataset, the topics identified by LDA can vary for different runs due to the use of random initialization when learning the model. To reduce the variation in results, we bias the word probabilities for each topic by initializing $\boldsymbol{\beta}$ in Blei et al. [4]. For an LDA model with k topics, $\boldsymbol{\beta}$ is a $k \times V$ matrix where $\beta_{ij} = p(w^j = 1 | z^i = 1)$, the probability of word w^j in topic z^i (the notation x^i denotes the i^{th} element of vector x). We first perform hard clustering on the documents in D, with the number of clusters set to the desired number of LDA topics, k. Then we initialize the values of β_i. for $i \in \{1, \ldots, k\}$ with the term distributions from hard clustering. That is, the k^{th} β_i. is initialized to the vector representing the normalized term distribution of the documents assigned to the k^{th} cluster. We chose to use spectral clustering because the clustering results are less variable.

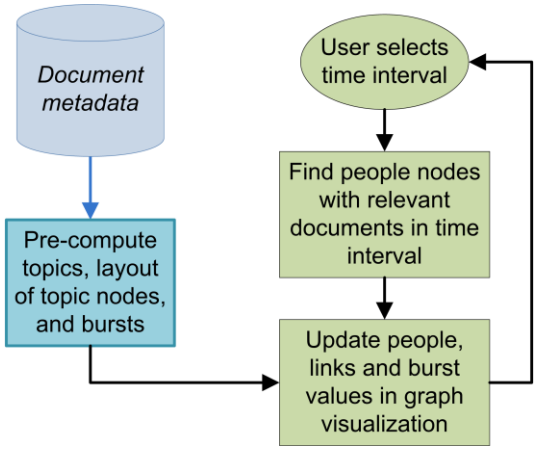


Figure 3. Overview of data analysis and interactive visualization process.

Dimension Reduction for Scalability

After the topics have been computed using LDA, we further reduce the dimensions for scalability and efficiency. Dimension reduction is performed on V using the LDA topics as follows. Each topic is composed of component terms with associated probabilities of the word conditioned on the topic. The probabilities are sorted and since the values decrease rapidly, we keep only m terms for each topic (we use $m = 10$). The component terms may overlap between the k topics. The union of these component terms creates a subspace V_C of V, with $\dim(V_C) \leq k*m$. By reducing the dimension, we can scale to large datasets. It also makes it more efficient in computation time and memory space. This is important if the data needs to be transmitted to Web browsers for the visualization application.

Layout of Topics

The topics in the visualization are represented by nodes of a graph. These Topic nodes are placed in fixed locations in the layout. To determine the Topic node coordinates for the visualization, we use multidimensional scaling (MDS), based on the normalized cosine similarity between topics in the vector space of terms. We compute the cosine similarity with respect to V_C.

Burst Detection for Topics

Burst detection is computed for the topics. For each topic, the number of relevant documents are found for each time point to generate an activity stream. A burst detection algorithm [13] is applied to the activity stream of each topic. Then for each time point of a topic, a burst level is

assigned: {0, 1, 2, 3}, where the highest activity level is labeled 3. These are mapped to different "heat" colors in the visualization: {green, yellow, orange, red}.

Interactive Visualization

The visualization is designed to interactively show the temporal dynamics of the people collaborating over topics. Similar to Collaboration Map [18], we use a node-link force directed graph with two types of nodes (fixed and moveable) and a time interval slider. For scalability, we employ Topic nodes instead of Publication Venue nodes, and the layout of the fixed nodes has been improved by using MDS.

To handle larger numbers of nodes and links, the interactive visualization is implemented in C++ with the Qt application framework [21] and runs on a desktop computer. To better visualize the nodes and their text labels, a high resolution 4K display can be used (Figure 1 and Figure 2b).

Timeline Widget

A timeline widget with a slider and a dial allows the user to select a point in time, set the size of the time interval, and drag it along a timeline (Figure 2a). When the selected time interval is changed, the visualization is updated to reflect the collaborations that occurred in the time interval.

Topic Nodes

The topics in the visualization are represented by square nodes in the graph. The Topic nodes are fixed in the positions obtained from MDS layout computation. Each Topic node is labeled with its top three component terms, i.e. the three terms with largest $p(w^j|z^i)$ along with an ID number.

The user can add a label to a Topic node. The component terms and the user's domain knowledge of the dataset may suggest a word or phrase for labeling a topic. Clicking on a Topic node pops up a widget for entering a topic label. The topic label is visualized in square brackets, replacing the topic ID number used as the default label. Figure 2b shows a topic label being entered in a widget, and the bottom-right of the figure shows a Topic node with a previously entered label "quantum computing".

The Topic nodes are colored based on the maximum burst level of the selected time interval. See Figure 2b. The size of the color patch represents the percent of activity during that time period with respect to the topic's total activity. As the time slider is moved, it is possible to see a topic's popularity level increase and then decrease, simultaneously with the other topics' changing popularity level.

Person Nodes

The Person nodes are transitory and can appear/disappear and move around as the time interval on the timeline widget is manipulated. The movement of the Person nodes is animated.

The selected time interval is used to filter the Person nodes: only the people with documents authored in this period are shown in the visualization. The links between Person nodes with co-author relationships in the selected time period are highlighted (in purple). Links are shown between Person nodes and their relevant Topic nodes (in gray). See Figure 1 and Figure 2b. These links define forces that drive the animation to layout the positions of the Person nodes. The Topic nodes are anchored in the MDS layout locations and do not move.

A Person node can be selected and dragged to see it and its links better. This is useful when there is clutter around the Person node. When the node is let go, it bounces back to its location in the force-directed layout.

A relevance score threshold parameter controls how many Person nodes are filtered and visualized. The nodes with scores above the threshold are retrieved. A Person node's score with respect to a Topic node is based on the cosine similarity of the person's documents to the topic. A document may be relevant to more than one topic. The relevance score can be set to achieve different purposes, such as finding highly relevant Person nodes by using a high value, or to reduce clutter and occlusion by using a moderate value.

Note that the documents are not rendered in the visualization, since the focus of the visualization is collaboration over topics. If desired, a simple feature can be used to get information about the documents; e.g. by showing a list of relevant documents when a Person node or Topic node is clicked.

EVALUATION

We performed evaluation of the computed LDA-based topics and our visualization using these topics to show collaborations among authors over time. The LDA-based topics were compared for similarity to the topics contestants identified as part of the InfoVis 2004 Contest [7]. Then we describe using the visualization to explore larger datasets.

Datasets

Our evaluations were performed using two datasets: (1) a public dataset from InfoVis Contest, (2) publication metadata from our lab.

The InfoVis dataset contains metadata for InfoVis papers spanning 8 years (1995-2002) plus their references, which we refer to as *IV8+R*. The core set of only InfoVis papers we refer to as *IV8*.

We created an extended dataset covering 20 years of InfoVis papers, which we refer to as *IV20*. This dataset is based a cleaned version of IV8+R produced by Ke et al. [11]. From this version we extracted a cleaned version of IV8, and then extended it by adding InfoVis papers from 2003-2014 using the openly accessible metadata in the IEEE digital library [9]. The number of InfoVis papers from IV8 to IV20 increased from 152 to 557, and the number of authors increased from 313 to 1075.

The dataset composed of publication metadata from our lab was extracted from an internal database. It contains metadata from 1995-2015 for 524 publication items by 306 authors over a period of 21 years. We refer to this dataset as *Lab21*.

Comparison of Topic Keyword Sets

We first compared the similarity of our LDA-based topics against both the manually and automatically defined topic keywords from the InfoVis 2004 Contest. There were six sets of keywords: two were produced manually and four were computed automatically. These keyword sets are summarized in Table 1.

The keywords were defined on different subsets of the InfoVis 2004 Contest dataset. Lee (ID 3) clustered only the InfoVis papers and not their references. Tyman (ID 5) used topics presented in an Information Visualization class. Keim (ID 2) manually defined topics after manual cleaning that included adding keywords to publications without keywords.

Ahmed (ID 0) removed papers that had no words remaining after preprocessing that included removing infrequent words, removing popular words such as "information" and "visualization", and then created a self-organizing map.

Teoh manually added keywords to documents without keywords and then automatically identified the most frequent keywords. Ke used a burst detection method on the full InfoVis dataset.

Since we performed our evaluation on the IV8 dataset, we identified keywords on IV8 based on Teoh's frequency method and refer to the keywords as "frequent" (ID 6), and on Ke's burst detection method and refer to the keywords as "burst" (ID 7).

Since only a few of the topic keywords overlap across documents, we use the word2vec representation [19] of the keywords when computing similarity. Specifically, we used the word and phrase vectors that were pre-trained on about 100 billion words from part of the Google News dataset [24]. This dataset contains about 3 million words and phrases represented as 300-dimensional vectors.

While a common approach to representing words in a document is to assign each word in the vocabulary an index in a term vector, word2vec learns an embedded representation of words, so that each word or phrase that word2vec was trained on has a reduced dimension, embedded representation. It has been shown that similar words are located in close proximity in the embedded representation [19]; thus, the word2vec representation allows a more fine-grained assessment of the similarity of topic keyword pairs.

To compare the sets of topic keywords from the InfoVis Contest and LDA, the normalized cosine similarity of pairs of topic keyword sets represented using word2vec as $\vec{x}$ and

ID	*Method*	*# Kwds*	*Manual*	*Dataset*	*Topic Keywords*
0	Ahmed [1]	11	no	IV8+R, filtered	Database Query and Data Mining, System Design, Web Data, Interaction, Graph Drawing, Focus + Context Techniques, Software Visualization, Hierarchy, Multidimensional Data Analysis, Trees, Text and Image Information Retrieval
1	Ke [11]	6	no	IV8+R	data visualization, focus + context, hierarchy, human factors, information visualization, user interface
2	Kiem [12]	5	yes	*N/A*	Information Visualization, HCI, Data Analysis, Computer Graphics, Graph Drawing
3	Lee [15]	5	no	IV8	General, Dynamic Queries, Graph Visualization, Focus + Context Techniques, Tree Visualization
4	Teoh [22]	15	partly	IV8+R, added keywords	text visualization, data mining, database, information analysis, information retrieval, multivariate data, taxonomy, user interfaces, data model, cartography, animation, internet, focus+context visualization, graph drawing, hierarchies
5	Tyman [23]	5	yes	*N/A*	Data, HCI, Implementation, Other, Techniques
6	frequent	15	no	IV8	data visualization, information visualization, world wide web (www), user interface, information retrieval, visualization, internet, focus+context, hierarchy, computer animation, interactive system, spreadsheet, graph drawing, animation, treemap
7	burst	6	no	IV8	information visualization, world wide web (www), user interface, information retrieval, visualization, graph drawing
8	LDA	5	no	IV8	[magnification, nonlinear, …], [document, text, …], [tree, hierarchy, …], [display, analysis, …], [query, database, …]

Table 1. Keyword sets by groups for the InfoVis 2004 Contest dataset. For LDA (ID 8), the two most probable keywords for each topic are shown in square brackets.

$\vec{y}$ were computed:

$$Sim(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|}$$

For each similarity computation, one keyword set is considered as the reference, and the other as the test. In part due to the varying number of keywords, the similarities were compared by computing the maximum similarity of each keyword in the reference against all the keywords in the test, and then the average of the keyword similarities in the reference was used as a measure of how well the test keywords align with the reference:

$$\overline{Sim(\vec{x})} = \frac{1}{|R|}\sum_{i\in R}\max_{j\in T} Sim\,(\vec{x_i},\vec{y_j}),$$

where R is the set of reference keywords and T is the set of test keywords.

A heat map indicating the similarity of each reference topic keyword set against each test topic keyword set is shown in Figure 4. Since three of the six InfoVis 2004 keyword sets had five topics, we used the LDA model with five topics for this comparison. We note that the similarity of the topic keyword sets is noticeably lower when the Tyman keywords (5) are the reference set. The Tyman keywords are general words which are not specific to the information visualization domain. We also note that when the LDA keywords are the reference, the similarity with the other keyword sets is also low, except for keyword set 0, Ahmed, which contains 11 keywords and possibly keyword set 3, Lee, who also employed clustering. Keyword set 4, Teoh, which contains the largest number of keywords, 15, also has a relatively high similarity with the other keyword sets. Of the frequency-based keyword sets computed from IV8, we note that both exhibit good similarity with the other InfoVis Contest keyword sets, and that the burst method (ID 6) is a subset of the frequency method (ID 7). In contrast, we observe that LDA, which models documents as a mixture of topics, produces keywords that are less similar on average to the other keyword sets.

We next examine how well each of the different keyword sets match the documents in the IV8 corpus. For this we compute the average normalized cosine similarity of a keyword set against each document. We again use the word2vec term representation. To create the document representation, the word2vec term vector for each of the words in a document is summed, similar to the way that the commonly-used term vectors representing the words in a document are summed to represent the document as a term-frequency vector. That is, the vector representation for a document, $\vec{d}$, is:

$$\vec{d} = \sum_{i\in W}\vec{v}_i,$$

where $\vec{v}_i$ is the word2vec representation for word i and W is the set of words in the document. When a keyword containing more than one term is not in the word2vec

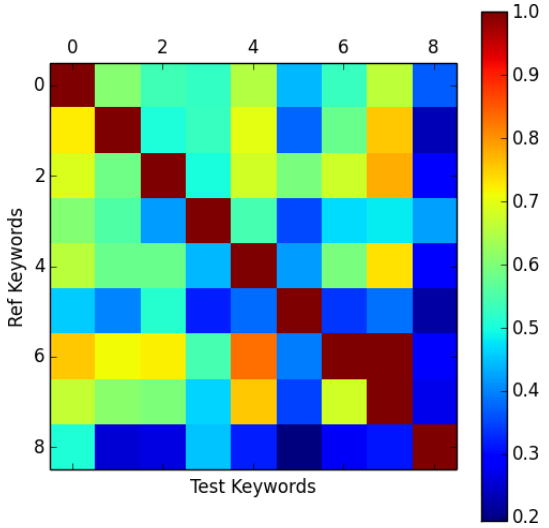


Figure 4. Heatmap showing the similarity of topic keyword sets. The numbers correspond to the ID for the different keyword identification methods shown in Table 1.

vocabulary we used, the word2vec representation of the component words in the phrase are summed and normalized to represent the phrase.

We represent LDA topics with one or more of the most probable terms or "topic keywords", i.e. the largest $p(w_i|z_j)$, over the words w_i in topic z_j. When more than one keyword is used to represent an LDA topic, the weighted average of the most probable N keywords is used to represent the topic, with weight $p(w_i|z_j)$ for each keyword w_i. The normalized cosine similarity of the document against each topic (represented as weighted keyword vector) in a keyword set is computed, and the similarity of the best-matching topic, i.e. the maximum document similarity, is selected as the similarity of the document to the keyword set.

The average document similarity for each keyword set is shown in Figure 5. The format of the names in the figure is [*identifier*]-[*info*]-[*#topics*]. For the InfoVis keywords, [*info*] specifies whether the keywords were manually defined or computed. For the LDA models, [*info*] indicates the number of most probable terms used in the similarity computation. We used {2, 3, 10} for comparison. Except for Tyman, the average number of keywords for all the InfoVis keyword sets is 2; the number of keywords shown in our visualization is 3, and 10 weighted keywords were used for similarity computations in our visualization.

From Figure 5 we observe that manually defined keyword sets performed the worse. The LDA keyword sets with two and three keywords per topic generally performed slightly worse than the InfoVis keyword sets in gold. Increasing the number of keywords per LDA topic to 10 improved performance in all cases to better than the InfoVis keyword sets in gold. Based on these results and the observation that the probability of the keywords associated with a topic

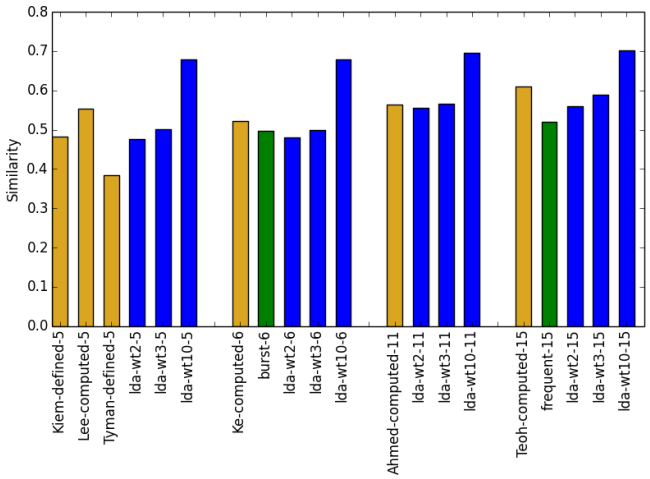


Figure 5. Similarity of keyword sets against documents in the IV8 dataset. The LDA keyword sets are in blue. The InfoVis 2004 Contest keywords are in gold. Frequency-based keywords computed from IV8 are shown in green.

which rank lower than the first ten is very small, we find support for using only the top ten keywords in placing the topics and when computing similarity between topics and documents.

Exploring Larger Datasets

We used our interactive visualization to explore the two larger datasets with longer time frames: IV20 and Lab21.

Example on InfoVis 20-year Dataset

As an example, we describe how the visualization shows a person changing his collaborators and topics over time. We ran the visualization on the InfoVis extended dataset IV20, and observed the person Jean-Daniel Fekete. To reduce clutter, the relevance threshold for Person nodes was set to 0.5. We set the time interval size to 3 years, and then we moved the time slider to examine different periods. Close-up regions of some interesting periods are shown in Figure 6. In the period 2002-2004, Fekete collaborated with two people on topic 4 (graph, layout, tree). In 2004-2006, Fekete continued to collaborate with the same two people, and also started collaborating with another person on topic 18 (network, social, structure). Topic 4 and 18 are somewhat similar as shown by their topic words (i.e. "graph" and "network") and their locations in the visualization. Later in 2011-2013, Fekete collaborated with a new group of people on topic 19 (analysis, display, knowledge), which is a different topic located farther in the visualization.

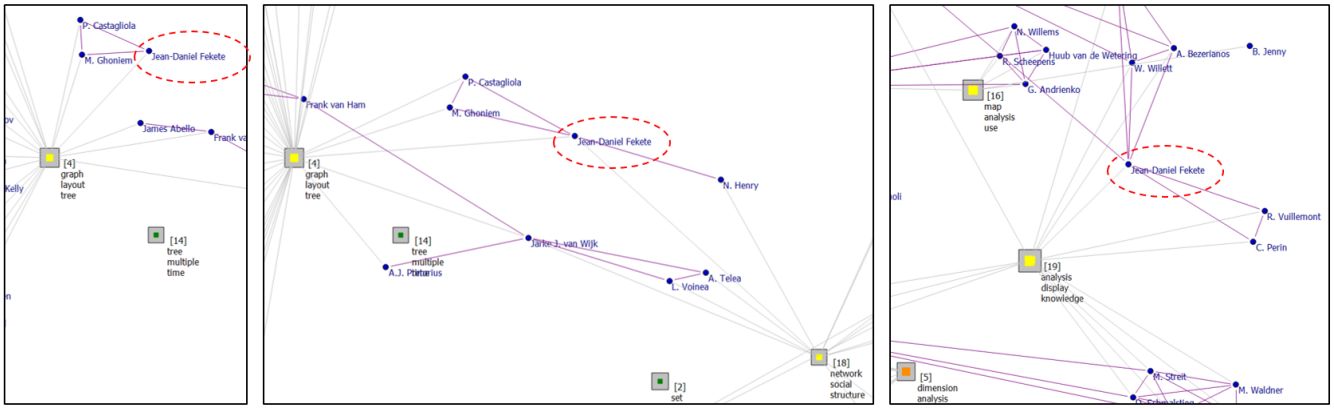
Observations on the Two Datasets

The dimension of the vector spaces V for the IV20 data is 8,252, and for the Lab21 data is 5,626. We use $m = 10$ for the number of highest probability components, so the dimension of subspace V_C subspace is $\leq 10k$, where k is the number of topics. In our tests with $k = 5, 10, 20$, the dimension of V is reduced by a factor of about 30 to 160.

We examine the sizes of the node-link graphs for different number of topics ($k = 5, 10, 20$), and for various relevance scores computed by cosine similarity.

The total number of Person nodes (*p-nodes*) and edges (person-person and person-topic) for different number of topics, *k,* are shown in Figure 7. The number of persons in the IV20 data is 1,075, and in the Lab21 data is 306. By looking at the p-nodes in Table 2, we see that the space of persons (and documents) is well covered. With IV20, for $k = 20$ all 1,075 p-nodes are covered, and for $k = 5$, only 2 p-nodes are missed. With Lab21, for $k = 20$, just 4 of the 302 p-nodes are missed. The missing nodes occur when the person's documents are not in the subspace V_C.

Having good coverage or recall of the p-nodes is important so that parts of the data do not become inaccessible in the visualization, and we see that the coverage is good by the topics and the subspace V_C under the dimension reduction. In contrast, a more simplistic approach such as taking the top k keywords [22] and using keyword matching would provide very poor coverage.



[2002-2004] [2004-2006] [2011-2013]

Figure 6. Time sequence with IV20 data. Close-up of regions in the visualization are shown for three different time periods.

The total number of nodes and edges can be large (see Table 2), which leads to clutter in the visualization. The nodes and edges can be filtered by varying the relevance score threshold so that only the more relevant objects are displayed. Figure 7 shows the number of p-nodes and edges retrieved with different relevance thresholds.

Dataset	p-nodes, $k = 5$	edges, $k = 5$	p-nodes, $k = 10$	edges, $k = 10$	p-nodes, $k = 20$	edges, $k = 20$
IV20	1073	3584	1075	3699	1075	3770
Lab21	292	1649	302	1759	302	1850

Table 2. Number of Person nodes (p-nodes) and edges.

The burst results show that there are roughly 5 to 7 bursts per topic over the 20 year periods (for $k = 5, 10, 20$). See Figure 8. For the IV20 data, level 2 bursts occur only at $k = 20$, and for the Lab21 data, level 3 bursts occur only at $k = 20$. One possible explanation is that the topics are more diverse in a research lab database with heterogeneous information than in conference paper metadata focused on a single research area, and that topics also become more diverse when there are more of them at higher k values.

CONCLUSION AND FUTURE WORK

We presented methods for interactive visualization of the temporal dynamics of collaborations over topics. Topic modeling was used to automatically identify topics to serve as landmarks in the visualization, MDS was used for placing the topics in the layout, and burst detection was applied to indicate active topics during a selected time period.

We compared LDA topic modeling against the topic keyword sets from the InfoVis Contest entries. Our results showed that the topic keywords selected by LDA differ more relative to the InfoVis topic keywords than the InfoVis topic keywords differ among each other. We observed that the additional terms provided by LDA-based keyword sets result in improved similarity between a topic keyword set and the documents in a corpus.

We used our topic-based method to visualize two larger datasets with longer time frames, one with only one publication venue, and a second with many venues. Our analysis demonstrated that the use of topic modeling allows for good coverage of the people in the visualization and the potential for scaling to larger datasets.

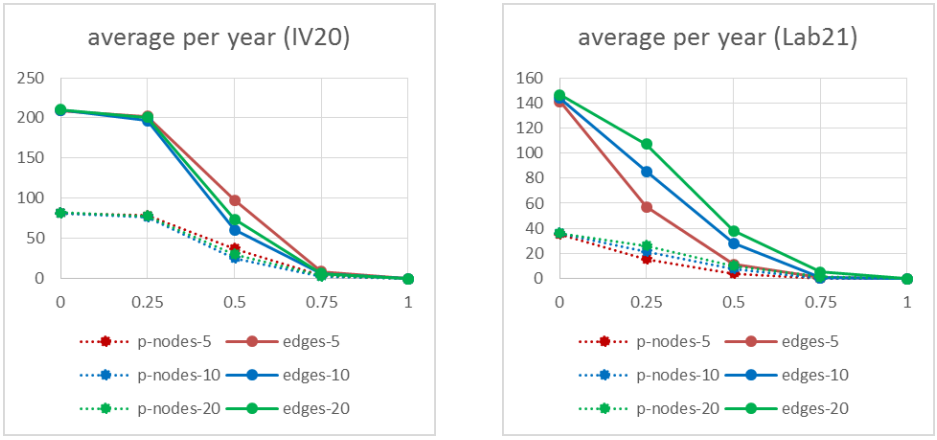


Figure 7. Number of Person nodes (p-nodes) and edges, as the relevance threshold is varied.

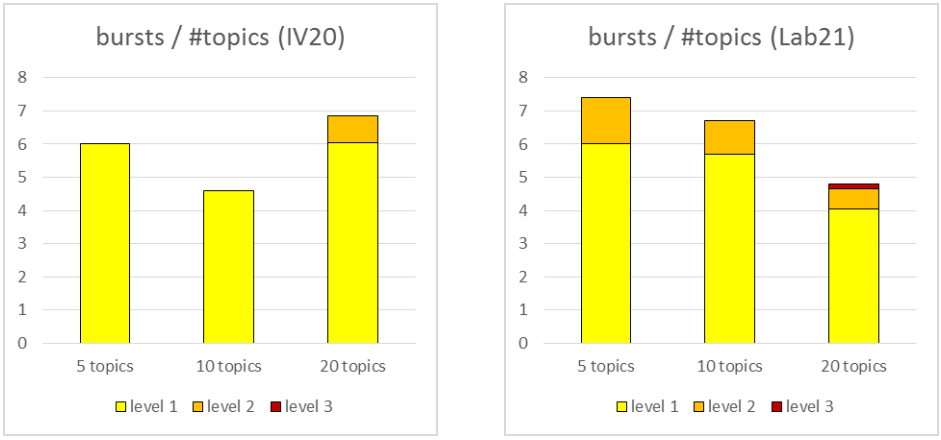


Figure 8. Burst results.

There are several directions in which our work could be improved and extended. These include adapting the visualization to smaller and lower resolution displays such as smartphones. With larger form factors such as wall-sized displays, more topics can be visualized; however, "junk topics" are more likely to occur. A good stop word set eliminates some of the junk topics, but others are due to sparse support [20] and require methods for identifying them (e.g. [2, 5]). Another possible extension is to explore whether alternative representations for the topic nodes may be more informative than the most probable terms that we used; for example, identifying phrases in addition to the single terms provided by LDA.

ACKNOWLEDGMENTS

We thank Bee Liew and Joel Tan for help with collecting the data.

REFERENCES

1. Ahmed, A., Dwyer, T., Murray, C., Song, L., Wu, Y.X. Wilmascope graph visualization. *IEEE InfoVis 2004 Contest Poster Compendium*.
2. AlSumait, L., Barbara, D., Gentle, J., Domeniconi, C. Topic significance ranking of LDA generative models. *Proc. ECML PKDD 2009*, pp. 67-72.
3. Arrow, H., Poole, M.S., Henry, K.B., Wheelan, S., Moreland, R. The temporal perspective on groups. *Small Group Research*, 35, 1 (2004): 73-105.
4. Blei, D.M., Ng, A.Y., Jordan, M.I. Latent Dirichlet allocation. *J. Machine Learning Research*, 3 (2003): pp. 993-1022.
5. Chuang, J., Manning, C. D., Heer, J. Termite: Visualization techniques for assessing textual topic models. *Proc. AVI '12*, pp. 74-77.
6. Faridani, S., Bitton, E., Ryokai, K., Goldberg, K. Opinion space: a scalable tool for browsing online comments. *Proc. CHI '10*, pp. 1175-1184.
7. Fekete, J.-D., Grinstein, G., Plaisant, C. IEEE InfoVis 2004 Contest, the history of InfoVis, www.cs.umd.edu/hcil/iv04contest.
8. Gensim Python software library. http://radimrehurek.com/gensim
9. IEEE Xplore Digital Library. http://ieeexplore.ieee.org/Xplore/home.jsp
10. Jacovi, M., Soroka, V., Gilboa-Freedman, G., Ur, S., Shahar, E., Marmasse, N. The chasms of CSCW: a citation graph analysis of the CSCW conference. *Proc. CSCW 2006*, pp. 91-101.
11. Ke, W., Borner, K., Viswanath, L. Major information visualization authors, papers and topics in the ACM Library. *IEEE InfoVis 2004 Contest Poster Compendium*.
12. Keim, D. A., Barro, H., Panse, C., Schneidewind, J., Sips, M. Exploring and visualizing the history of InfoVis. *IEEE InfoVis 2004 Contest Poster Compendium.*
13. Kleinberg, J. Bursty and hierarchical structure in streams. *Proc. KDD 2006*, pp. 289-298.
14. Kraker, P., Weißensteiner, P., Brusilowsky, P. Altmetrics-based visualizations depicting the evolution of a knowledge domain. *Proc. STI 2014*, pp. 330-333.
15. Lee, B., Czerwinski, M., Robertson, G., Bederson, B.B. Understanding research trends in conferences using PaperLens. *Proc. CHI 2005*, pp. 1969-1972.
16. Leskovec, J., Kleinberg, J., Faloutsos, C. Graphs over time: densification laws, shrinking diameters and possible explanations. *Proc. KDD 2005*, pp. 177-187.
17. Ley, M. Digital bibliography & library project (DBLP). http://www.informatik.uni-trier.de/~ley/db/.
18. Lim, S., Chiu, P. Collaboration Map: Visualizing temporal dynamics of small group collaboration. *CSCW 2015 Companion*, pp. 41-44.
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. Distributed representations of words and phrases and their compositionality. *Proc. NIPS 2013*, pp. 3111-3119.
20. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A. Optimizing semantic coherence in topic models. *Proc. EMNLP 2011*, pp. 262-272.
21. Qt software application framework. https://www.qt.io/developers.
22. Teoh, S.T., Ma, K.-L. One-For-All: Visualization of the information visualization symposia. *IEEE InfoVis 2004 Contest Poster Compendium.*
23. Tyman, J., Gruetzmacher, G.P., Stasko, J. InfoVisExplorer. *IEEE InfoVis 2004 Contest Poster Compendium*.
24. Word2vec. https://code.google.com/p/word2vec/
25. Zhao J., Collins C., Chevalier F., Balakrishnan R. Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Trans. on Visualization and Computer Graphics,* 19, 12 (2013): 2080-2089.