Spatiotemporal Analytics of Topic Trajectory

Jiangen He Drexel University
3141 Chestnut Street
Philadelphia, Pennsylvania, USA
jiangen.he@drexel.edu

Chaomei Chen Drexel University
3141 Chestnut Street
Philadelphia, Pennsylvania, USA
chaomei.chen@drexel.edu

ABSTRACT

Spatially and temporally relevant text data generated on the Internet by users worldwide is of great value for investigating and understanding emerging trends of user interests and how they may evolve over time and space. However, exploring the spatiotemporal text data and characterizing the evolution of topics over time and space are challenging due to the complexity of such data and associated activities. This paper proposes a new approach to exploring the spatiotemporal text data with visual filters. We introduce a notion of topic trajectory to depict the spatiotemporal evolution of topics. Multiple coordinated visualizations provided in our visualization system enable users to explore topic trajectories and develop their contextual awareness in terms of how information flows across different regions. We demonstrate the use of our system with an analysis of a dataset contributed by users of widely used science mapping software CiteSpace.

CCS Concepts

•**Human-centered computing → Visual analytics; Geographic visualization;**

Keywords

Topic Trajectory; Temporal Visualization; Spatial Visualization; Visual Analytics

1. INTRODUCTION

Large volumes and varieties of time-stamped and geospatial data generated by users on the Internet are increasingly available due to advances in technology and the popularity of social media, such as Twitter, Facebook, and Youtube. Prior to the age of social media, the datasets containing IP addresses can be indexed as spatial data. Spatiotemporal data is very common in real-world settings. However, as the amount of data has exceeded the capabilities of manual evaluation, especially when dealing with a large amount of text, there is a need for advanced techniques to aid our understanding of large volumes of spatiotemporal text data.

Numerous text analysis technologies have been developed to uncover hidden thematic structures in text collections [1, 2], but results of these text analysis technologies are often complex and not straightforward for users to understand. To aid users to develop a full picture of a text collection through text analysis, many advanced technologies couple the state-of-the-art text analytics with interactive visualization [3, 4]. Studies of topic visualization can facilitate the understanding and analysis of text collections based on topics. However, few efforts allow users to explore the topics of a text collection from an integrated spatiotemporal perspective and depict a full picture of how topics evolve over time and space.

Exploring topics of a text collection over time and space can aid users to identify the spatiotemporal change of a topic, identify emerging trends of information diffusion and understand the information interaction between geographical regions. We, therefore, contribute a spatiotemporal visual analytics system to the study of such topical dynamics by depicting the spatiotemporal evolution of a topic as a trajectory on a geographical map and providing multiple interactive and coordinated views [5] to support an interactive topic analysis. As a result, we offer these contributions:

- **A visual analytics system** that facilitates users to explore and analyze the evolution of topics over time and space.
- **A notion of topic trajectory** which provides an intuitive approach that aids users to understand the evolution of a topic and simplifies a large collection of trajectories so as to improve the clarity of an overview of trajectories.
- **Flexible visual interaction techniques** which enable users to visually filter the topics and regions during an exploration through multiple coordinated views.

2. RELATED WORK

2.1 Topic Visualization

The technologies of topic visualization are typically integrated with advanced text analysis technologies to take the advantages of both visual analytics and data mining. The existing topic visualizations focus on analyzing different aspects of the topics of a text collection. Many researchers

VINCI '16, September 24-26, 2016, Dallas, TX, USA
© 2016 ACM. ISBN 978-1-4503-4149-3/16/09. . . $15.00
DOI: http://dx.doi.org/10.1145/2968220.2968244

focus on depicting evolving topics over time. Topic visualizations in another category emphasize static relations among topics to present a full picture of text collection.

ThemeRiver [6] is a pioneering example of visualizing thematic flows over time. ThemeRiver has inspired many further applications on topic visualization. ParallelTopics [7], for example, illustrates topic evolution over time, but it also combines parallel coordinate plots with the probabilistic distribution of a document on different topics. TIARA [3] tightly integrates the stacked graph visualization and tag clouds with Latent Dirichlet Allocation [1] model to illustrate topic evolution patterns over time. Sankey diagrams is another commonly seen design option for topic visualization. TextFlow [8] and RoseRiver [9] leverage Sankey diagrams to visually convey topic merging and splitting relationships over time. Visualizations that emphasized static relations among topics often focus more on the overlaps among topics and the structures of topics, since without depicting the temporal evolution of topics, they can depict more complex structure of topics. HierarchicalTopics [10] hierarchically organizes the extracted topics by the BRT model [11] and thus can represent a large number of topics without being cluttered. FacetAtlas [12] adopts the density-based graph visualization to represent the multifaceted relationships of documents within or across clusters of documents. Although these visualizations depict evolving patterns or relations of topics, the relevant geographic context is not adequately represented in these visualizations.

2.2 Spatiotemporal Visual Text Analysis

Although text data with spatiotemporal information is common, visualization studies are still relatively few. Since spatiotemporal text data has numerous features to be visualized, including the topics, time, and space, multiple coordinated views provide an ideal platform [5]. VisGets [13] allows users to formulate queries that simultaneously combine temporal, spatial, and topical data filters by providing visual overviews and offer visual filters. ScatterBlogs2 [14] suggests a new approach such that analysts can build task-specific message filters interactively based on recorded messages of well-understood previous events. Whisper [15] highlights three primary characteristics of diffusion processes in social media: temporal trends, social-spatial extent, and community responses to a topic of interest. These studies provide exploratory visualizations of topics with a geographic context, but investigating evolving pattern of numerous topics and the role of a geographical region in a topic evolution is not supported.

3. PROBLEM

3.1 Data

The data we used to drive our initial design of this system is from the log data of CiteSpace. CiteSpace [16, 17] is a freely available visual analytics application for users to analyze and visualize a knowledge domain based on relevant articles published in the literature. In essence, CiteSpace detects and visualizes grouping patterns of the topics of a set of scientific publications based on a series of the citation networks. This method is generic and applicable to a wide variety of scientific fields. CiteSpace has been used by users from many countries worldwide. With consents of users, CiteSpace registers over 80 types of interactive events generated by users during a visual analytics session. We retrieved the geographic information of IP addresses to index the log events with geographic locations. After cleaning the data, our dataset consists of 35,446 records, i.e. events. Although this study is motivated by the log analysis of CiteSpace [18], our long-term goal is to design a visual analytics system which can also analyze other text collections with temporal and spatial information. For example, the hashtag dataset from Twitter can be analyzed directly with our system.

3.2 Task Analysis

The data of this study contains three primary types, namely time, space, and topic. In a typical visual analytics session with CiteSpace, the user would generate clusters of clusters with labels that characterize the topic of each cluster. The topic recorded in the log event corresponds to the largest component of the visualized network. Our goal is to design a system which can assist users to investigate such data across multiple dimensions.

T.1 What are the general patterns of the evolution of topics? Geographic factors might have effects on user interests or the chances they receive information. Users may introduce new topics in a geographic region, adopt topics originated from some specific regions, or ignore some of the topics. There are many possible general patterns one may learn from the overview of topic trajectories.

T.2 Which regions are involved in the evolution of a topic or a group of topics and to what extent? Regional contributions to a topic or a group of topics may vary greatly across regions. When investigating a topic, the system should provide contextual information to enable users to explore the role of each region and the degree of its involvement.

T.3 How does information flow across different regions? Topic trajectories indicate patterns of information flow to an extent. Interactivity is useful to allow users to investigate the topic trajectories of their interests.

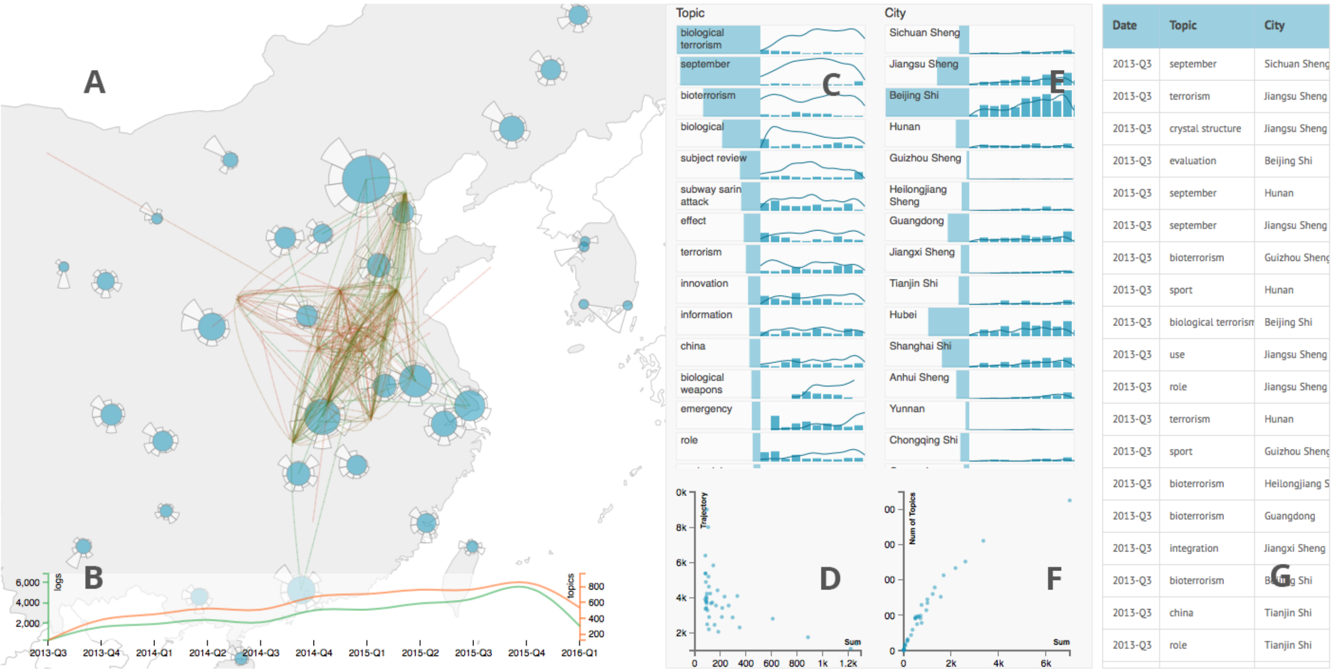
4. VISUALIZATION DESIGN

4.1 System Overview

Figure 1 presents an overview of our visualization system. The main interface consists of three views: a map view (A), a list view (C, E, G, D, and F), and a time view (B). All these views are coordinated. The map view (A) includes a glyph as a node to represent a region and topic trajectories to show the evolution of topics. In the list view, the lists of topics (C), cities (E) and logs (G) and scatterplots of topics (D) and cities (F) are shown. The time view (B) provides an overview of logs with a time filter. In addition, our system supports various interactions, such as filtering and selecting in different views. For example, upon selecting a topic from the list view, the corresponding topic trajectory will be highlighted, the region glyph will show the number of logs containing this topic over time and related logs are also highlighted in the log list view. Combining all the views enables users to analyze topics at different levels of granularity.

4.2 Visual Design

Topic Trajectory The geographic position of a topic at a time point is computed as the weighted centroid of all the regions that contain log events associated with the topic. A similar design was proposed in [19]. The position of a topic



Date	Topic	City
2013-Q3	september	Sichuan Sheng
2013-Q3	terrorism	Jiangsu Sheng
2013-Q3	crystal structure	Jiangsu Sheng
2013-Q3	evaluation	Beijing Shi
2013-Q3	september	Hunan
2013-Q3	september	Jiangsu Sheng
2013-Q3	bioterrorism	Guizhou Sheng
2013-Q3	sport	Hunan
2013-Q3	biological terrorism	Beijing Shi
2013-Q3	use	Jiangsu Sheng
2013-Q3	role	Jiangsu Sheng
2013-Q3	terrorism	Hunan
2013-Q3	sport	Guizhou Sheng
2013-Q3	bioterrorism	Heilongjiang S
2013-Q3	bioterrorism	Guangdong
2013-Q3	integration	Jiangxi Sheng
2013-Q3	bioterrorism	Beijing Shi
2013-Q3	china	Tianjin Shi
2013-Q3	role	Tianjin Shi

Figure 1: System Overview. (A) A map view of topic trajectories and regional nodes. (B) A timeline view of logs and topics over time. Time filtering is supported by this chart. (C) A list view of topics of users' interest. (D) A scatterplot of topics. (E) A list of topic originating regions (Sheng=Province in Chinese, Shi=City in Chinese). (F) A scatterplot of cities (G) A log list. All the views are interactively coordinated.

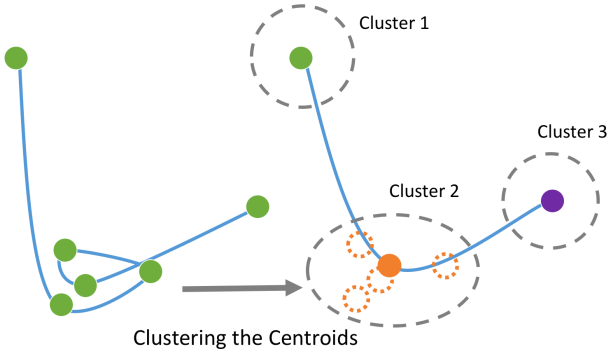


Figure 2: Simplifying a topic trajectory: nodes on trajectories are clustered by their proximities and nodes in the same cluster (cluster 2) will be merged into a new single one (the solid one in the cluster 2) located at the cluster centroid.

may change over time because of the change of the spatial distribution of log events associated with the topic. The trajectory of topic positions over time represents the evolution of the topic in space. A curve connects topic positions at adjacent time points. A point on a topic trajectory summarizes the spatial distribution of the topic. A topic trajectory thus depicts the spatiotemporal change of the topic. Each segment of a trajectory is color coded, starting in green and ending in red such that the greener part of a specific trajectory depicts the spatiotemporal change of the topic at earlier stages.

However, in many scenarios, the position of the centroid of a topic may remain relatively stable with minor fluctuations only, and different topics may share a similar evolution pattern. Minor fluctuations of a specific topic and minor differences among similar topics are less important information and they may reduce overall clarity of an overview showing numerous topic trajectories. To increase the saliency of important patterns, we propose a design that simplifies these trajectories. We group the centroids into clusters and then merge all the centroids in each cluster into a new single-point supercentroid. By connecting the merged nodes with other nodes, trajectories become simpler and smoother. In this way, topic trajectories and the overview of trajectories become more readable. The process of simplifying a trajectory is illustrated in Figure 2.

Glyph Design The nodes of regions in the map provide essential contextual information for topic trajectories as well as essential information on regions. To reveal the important features of each region, we proposed a glyph design to help users to identify the differences among regions and provide comprehensive contextual information for topic trajectories. Prior to selecting any topic, the glyph presents the basic information of regions. The diameter of the node represents the number of log events originated by users from this region. The pie chart slices around the node represents the number of events per time unit (per quarter in this case). When selecting one or more topics, the node and pie slice act as containers to show the contextual information for selected topic(s). The contextual information includes the number of log events containing the selected topic(s) and the number of these log events in each time unit. To show contextual information on multiple topics at the same time, all the information uses the unique color assigned when the topic is selected. The design of the glyph is shown as Figure 3.

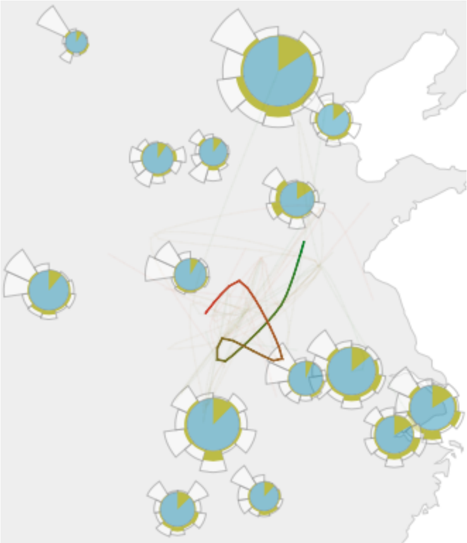


Figure 3: Glyph design and contextual information of a selected topic. A topic trajectory is selected and each region node shows the overall and temporal proportion of relevant logs in the region.

List View The list view includes a topic list, a region list, a topic scatterplot, a city scatterplot and a log list. The topic list and region list offer information-rich visual items to facilitate users to obtain a good understanding quickly. These two list views enable users to select topics and regions individually. Multiple selections are also allowed in these two lists. Aiming to facilitate users to select a group of items according to their attribute similarity, we design two scatterplots to allow group selection. The log list would be sorted according to the relevance of the selected items when new item selected.

The topic list reveals the information of the number of log (the horizontal bar), the movement of centroids in the topic trajectory (the bar chart) and the change of distributed regions over time (the line chart); The region list reveals the number of log (the horizontal bar), the change of involved topics (the bar chart) and the number of newly produced logs over time (the line chart).

4.3 Interactive Exploration

The interaction among the three views is carefully designed to support the exploration of topic data from different aspects and at various levels of detail. Both of individual and multiple selections of region(s) and topic(s) are enabled and visual selection of time is also supported. Topics and regions are two main types of entities in this visualization system, and we support interactive exploration for these two types of entities. When topics are selected, the corresponding topic trajectories will be highlighted and other topic trajectories will be faded with low opacity. As mentioned in Section 4.2, the regional glyph nodes show the context of selected topic(s) (see Figure 3). When one or more regions are selected, the corresponding topic trajectories will be highlighted, and opacity of topic trajectories will be adjusted according to the relevance to the selected region(s).

5. CASE STUDY

We analyze a dataset of interactive event logs from users of CiteSpace worldwide. The resultant visualization is shown in Figure 1. Given that East Asia is the most active area in the log records, we focus on this area in our case study. The visual analysis tasks specified in Section 3.2 were supported in our system.

T.1: The overview of the topic trajectories reveals the general pattern. Many green trajectories started from Beijing and some regions in southern China, which means that users from Beijing and some regions in southern China used CiteSpace to analyze a lot of new topics. Most of the red lines converged toward the center of China, suggesting relatively stable activities in this area. Most of the topics have spread to numerous regions.

T.2: Upon selecting the topics of interest, each node of a region depicts temporal information of the topics in this region so that we can accurately identify every change in the topic trajectory. We selected a topic in Figure 3, the pie chart in a region node shows the overall proportion of logs containing the selected topic in the corresponding region, and the color block in each pie slice around the node shows the proportion of relevant logs in a specific period.

T.3: Flow patterns of selected regional nodes have that the trajectories between Hubei, Guangdong, and Beijing have higher densities than the average. By selecting region nodes, trajectories with lower relevance will fade out to make the trajectories of our interest more salient.

6. DISCUSSION AND CONCLUSION

This visualization system is designed to help analysts to explore and understand various patterns of topic trajectories in a multidimensional spatiotemporal space. Three coordinated views, namely, the map view, the list view and the time view, are designed and implemented for analyzing topic evolution patterns from different perspectives. A preliminary analysis of a real-world dataset with this system was encouraging.

However, the system still has several limitations. The dataset conducted in this study is a well-extracted and well-structured data, and the original text is not analyzed in this system. This may have several implications: firstly, contextual information of topics provided in this system is currently limited; Secondly, relations between topics cannot be extracted and depicted in more detail in this visualization system; Thirdly, the clusters of topics cannot be detected in this system and because of this, these limitations need to be resolved when we move to another level of granularity.

Although the approach to simplify and smooth trajectories is provided in this study, the overview still contains many overlapping trajectories. Existing edge bundling techniques in graph visualization [20] and parallel coordinates [21] may offer useful insights. However, current techniques are not readily applicable here because they are mainly designed to bundle straight lines. Bundling curves remains a challenging issue that we plan to address in future work.

Our visual analytics system can be further improved. For example, we will apply the original text data and integrate text mining techniques to enhance the contextual awareness of users in exploring the text data. We also plan to optimize the trajectory overview and develop curve bundling techniques in our visualization.

7. REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[2] Mark Dredze, Hanna Wallach, Danny Puller, and Fernando Pereira. Generating summary keywords for emails using topics. *Proceedings of the 13th international conference on Intelligent user interfaces SE - IUI '08*, pages 199–206, 2008.

[3] Shixia Liu, Michelle X. Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. *ACM Transactions on Intelligent Systems and Technology*, 3(2):1–28, 2012.

[4] Shixia Liu, Xiting Wang, Jianfei Chen, Jim Zhu, and Baining Guo. TopicPanorama: A full picture of relevant topics. *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings*, pages 183–192, 2015.

[5] Jonathan C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, CMV '07, pages 61–71, Washington, DC, USA, 2007. IEEE Computer Society.

[6] S Havre, B Hetzler, and L Nowell. ThemeRiver: visualizing theme changes over time. *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, 2000:115–123, 2000.

[7] Wenwen Dou, Xiaoyu Wang, Remco Chang, and William Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 231–240, 2011.

[8] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, and Yangqiu Song. TextFlow: Towards Better Understanding of Evolving Topics in Text. 17(12):2412–2421, 2011.

[9] Weiwei Cui, Shixia Liu, Zhuofeng Wu, and Hao Wei. How hierarchical topics evolve in large text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2281–2290, 2014.

[10] Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2002–2011, 2013.

[11] C Blundell, YW Teh, and KA Heller. Discovering non-binary hierarchical structures with Bayesian rose trees. *Mixture Estimation and Applications*, 2011.

[12] Nan Cao, Jimeng Sun, Yu Ru Lin, David Gotz, Shixia Liu, and Huamin Qu. FacetAtlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1172–1181, 2010.

[13] Marian Dörk, Sheelagh Carpendale, Christopher Collins, and Carey Williamson. VisGets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1205–1212, 2008.

[14] Harald Bosch, Dennis Thom, Florian Heimerl, Edwin Puttmann, Steffen Koch, Robert Kruger, Michael Worner, and Thomas Ertl. ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2022–2031, 2013.

[15] Nan Cao, Yu Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2649–2658, 2012.

[16] Chaomei Chen. Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5303–5310, 2004.

[17] Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006.

[18] Qing Ping, Jiangen He, and Chaomei Chen. How Many Ways to Use CiteSpace? A Study of User Interactive Events over 14 Months. *Journal of the Association for Information Science and Technology.* (in press).

[19] Chaomei Chen and Loet Leydesdorff. Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the association for information science and technology*, 65(2):334–351, 2014.

[20] O. Ersoy, C. Hurter, Fernando V. Paulovich, G. Cantareiro, and A. Telea. Skeleton-Based Edge Bundling for Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2364–2373, dec 2011.

[21] Gregorio Palmas, Myroslav Bachynskyi, Antti Oulasvirta, Hans Peter Seidel, and Tino Weinkauf. An edge-bundling layout for interactive parallel coordinates. *IEEE Pacific Visualization Symposium*, pages 57–64, 2014.