Visual Exploration and Comparative Analytics of Multidimensional Data Sets

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Xiaotong Liu, B.S.

Graduate Program in Computer Science and Engineering

The Ohio State University

2016

Dissertation Committee:

Han-Wei Shen, Advisor

Yusu Wang

Arnab Nandi

Abstract

Recently, rapidly growing amounts of data with numerous attributes and variables arise in various areas of science, engineering, business, and others. Analysis of the multi-faceted information contained in multidimensional data sets has already led to breakthroughs in many fields and emergence of new information-based industries. For instance, insights from business data can enable marketing experts to make better decisions, such as optimizing operations, preventing threats and fraud, capitalizing on new sources of revenue, and deepening customer engagement. Data with high dimensionality and complexity has far exceeded our human ability for comprehension without powerful tools. While many advanced computational models and algorithms have been developed for discovering useful information in multidimensional data sets, these fully-automated approaches may not work well when users only have vague hypotheses or even no hypothesis before analyzing the data.

Visualization enhances human's understanding by organizing information in graphical display, offering the possibility of visual exploration of data for knowledge discovery and sense-making. Visual exploration strengthens human perceptual capabilities with visual interfaces to guide data navigation, actively engaging users into the exploration process to make knowledge discovery much more efficient. However, due to the increasing heterogeneity and complexity of multidimensional data, the multidimensional data space exceeds human comprehension. Novel representations are needed to display and organize data

items based on the relationships of the dimensions in multidimensional data sets. Furthermore, visual analysis of multidimensional data sets often requires investigating the hidden relationships between different dimensions and specific items to understand the multi-faceted properties of the data sets. The enormous multidimensional data space complicates the search of potentially interesting relations between dimensions and data items. Powerful and versatile visualization tools are thus needed to allow users to analyze and compare complex relations and heterogeneous structures in multidimensional data for knowledge discovery and sense-making.

In this dissertation, we investigate critical aspects of multidimensional data visualization and comparative analytics in assisting users in visual exploration of multidimensional data for knowledge discovery and sense-making. Specifically, we address the questions: *How can we design multidimensional visualizations to enable effective visual summarization of multi-faceted data characteristics? How can we enhance the power of multidimensional visualizations with comparative analytics to allow visual identification and explanation of complex data relationships?* Based on theoretical foundations of visualization and visual analytics, we approach the questions with various research methodologies such as creative design methods, quantitative studies and qualitative studies.

This dissertation contributes novel comparative visualization techniques for multidimensional data, novel multidimensional data analysis methods, design implications and guidelines, and generalizable visual exploration frameworks. First, we propose the CorrelatedMultiples visualization that organizes multidimensional data items in 2D display space based on their correlations and similarities in the original high-dimensional space. Second, we investigate the effects of representations and juxtaposition designs on graphical perception of adjacency matrix visualizations, and present the TileMatrix visualization

for visualizing a large number of matrices. Third, we describe a novel association analysis method to identify informative and unique scalar values in multivariate scientific data sets, and the visualization framework with multiple interactive views to explore the scalars of interest with confident associations in the multivariate spatial domain. Finally, we present the design study of the SocialBrands tool to assess and analyze public perceptions of brands on social media. This dissertation ends with important reflections on designing comparative visualizations for multidimensional data and discussions of future work.

This is dedicated to my parents Ruiping Jie and Yuecheng Liu

Acknowledgments

First of all, I would like to express my sincere gratitude to my advisor Dr. Han-Wei Shen for his continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. I was very lucky to work with him on various fundamental research topics in both information visualization and scientific visualization. His guidance helped me tremendously in all the time of research and writing of many research papers, which result in the completion of this dissertation. I could not have accomplished so much without his support and advice.

I must also thank Dr. Yusu Wang and Dr. Arnab Nandi for contributing their valuable insights to various parts of my research, besides serving in my dissertation committee. I would like to thank Dr. Huamin Wang for kindly serving as the department representative in my candidacy examination, and thank Dr. Sebastian Kurtek for kindly serving as the graduate faculty representative in my dissertation examination. I am grateful to Dr. Srinivasan Parthasarathy for opening my eyes on research related to data mining and social computing.

My sincere thanks also goes to my mentors, managers and colleagues during my internship, who provided me opportunities to join their team as intern, and work on exciting projects in industry. A huge thanks to Yifan Hu and Stephen North who taught me a lot at the early stage of my research in information visualization. A huge thanks to Anbang Xu,

Liang Gou, Yi Wang, Haibin Liu and Rama Akkiraju who broadened my vision of research in industry.

I thank my fellow lab mates and friends Chun-Ming Chen, Wenbin He, Souyma Dutta, Ayan Biswas and Junpeng Wang for the productive collaborations and for the days and nights we were working together before deadlines. I cherish my memories of all the stimulating discussions and the fun I had with them and others in the *GRAVITY* group, including Xin Tong, Kewei Lu, Cheng Li, Tzu-Husan Wei and Ko-Chih Wang in the last five years. I am also grateful to the senior lab mates Teng-Yok Lee and Abon Choudhuri for their advice in my start as a Ph.D. student.

I would like to express my special appreciation to my family back in China. A huge thanks to my parents Ruiping Jie and Yuecheng Liu, my grandparents Xixiang Teng and Mingjun Jie, and my aunt Jie Jie. Their unconditional love and care have been the greatest support of my Ph.D. study. I love them so much, and I would not have made it this far without them. I especially thank my uncle Daniel Fertig and his mother Ellen Fertig, who always welcome me every time I visit New York.

The best outcome from these past five years in Columbus is finding my best friend and soul-mate Fang Liu, who has been a constant supporter and a great source of inspiration. We share so many interests, tastes, and opinions of life. I truly thank Fang for always accompanying me during my good and bad times. Our shared experiences over these years strengthened our commitment and determination to each other.

Vita

June 21, 1988 ..... Born - Yantai, China

2011 ..... B.S. Software Engineering,
Shanghai Jiao Tong University

November 2010 - May 2011 ..... Research Intern,
Microsoft Research Asia

2011 - 2012 ..... Graduate Teaching Associate,
The Ohio State University

June - August, 2012 ..... Research Intern,
AT&T Shannon Research Lab

May - August, 2015 ..... Research Intern,
IBM T.J. Watson Research Center

2011-present ..... Graduate Research Associate,
The Ohio State University

Publications

Research Publications

Xiaotong Liu, Anbang Xu, Liang Gou, Haibin Liu, Rama Akkiraju, and Han-Wei Shen, "SocialBrands: Visual Analysis of Public Perceptions of Brands on Social Media". *IEEE Visual Analytics Science and Technology*, Baltimore, Maryland, 2016.

Xiaotong Liu and Han-Wei Shen, "Association Analysis for Visual Exploration of Multivariate Scientific Data Sets". *IEEE Transactions on Visualization and Computer Graphics*, vol.22, no.1, pp. 955-964, January 2016.

Junpeng Wang, Xiaotong Liu, Han-Wei Shen, and Guang Lin, "Multi-Resolution Climate Ensemble Parameter Analysis with Nested Parallel Coordinates Plots". *IEEE Transactions on Visualization and Computer Graphics*, 2016.

Chun-Ming Chen, Soumya Dutta, Xiaotong Liu, Gregory Heinlein, Han-Wei Shen, and Jenping Chen,"Visualization and Analysis of Rotating Stall for Transonic Jet Engine Simulation". *IEEE Transactions on Visualization and Computer Graphics*, vol.22, no.1, pp. 847-856, January 2016.

Ayan Biswas, Guang Lin, Xiaotong Liu, and Han-Wei Shen, "Visualization of Time-Varying Weather Ensembles Across Multiple Resolutions". *IEEE Transactions on Visualization and Computer Graphics*, 2016.

Wenbin He, Chun-Ming Chen, Xiaotong Liu, and Han-Wei Shen, "A Bayesian Approach for Probabilistic Streamline Computation in Uncertain Flows". In *IEEE Pacific Visualization Symposium*, Taiwan, April 2016.

Xiaotong Liu and Han-Wei Shen, "The Effects of Representation and Juxtaposition on Graphical Perception of Matrix Visualization". In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 269-278, Seoul, Republic of Korea, April 2015.

Xiaotong Liu, Srinivasan Parthasarathy, Han-Wei Shen, and Yifan Hu, "GalaxyExplorer: Influence-Driven Visual Exploration of Context-Specific Social Media Interactions". In *International World Wide Web Conference*, pp. 215-218, Florence, Itay, May 2015.

Xiaotong Liu, Yifan Hu, Stephen North, and Han-Wei Shen, "CorrelatedMultiples: Spatially Coherent Small Multiples with Constrained Multidimensional Scaling". *Computer Graphics Forum*, January 2015.

Xiaotong Liu, Han-Wei Shen, and Yifan Hu, "Supporting Multifaceted Viewing of Word Clouds with Focus+Context Display".*Information Visualization*, vol.14, no.2, pp. 168-180, April 2015.

Xiaotong Liu, Yifan Hu, Stephen North, and Han-Wei Shen,"CompactMap: A Mental Map Preserving Visual Interface for Streaming Text Data". In *IEEE International Conference on Big Data*, pp. 48-55, Silicon Valley, CA, USA, October 2013.

Fields of Study

Major Field: Computer Science and Engineering

Studies in:

- Information Visualization
- Scientific Visualization
- Visual Analytics
- Human Computer Interaction

Table of Contents

Page

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Motivation

In the era of information explosion, the fast development of powerful data collection mechanisms and storage tools make it practical to collect an ever increasing amount of data from business, science, engineering, medicine, and many other areas. Analysis of the information contained in large-scale data sets has already led to major breakthroughs in fields of astronomy and climatology as well as new information-based industries. For example, businesses produce massive data sets, such as sales transactions, stock trading records, sales promotions, brand performance measures, and customer feedback. Insights from these data can enable marketing experts to make better decisions, such as optimizing operations, preventing threats and fraud, capitalizing on new sources of revenue, and deepening customer engagement. The prevalence of big data offers numerous opportunities that combine knowledge of preferences and needs with fine-grained information for decision making at the level of single individuals [23].

More often than not, the data to be analyzed are often rich in *dimensions* (also known as *variables*, *attributes*, or *facets*). For example, a brand performance data set can contain many attributes such as market share, purchase rate, purchase frequency, and portfolio size [30]. The challenges brought by massive multidimensional data go beyond storage,

indexing, and querying, which have been the province of classical data management research. Instead, they hinge on the ambitious goal of *inference* that turns multidimensional data into multi-faceted knowledge that are not explicitly present in the raw data. Data with high dimensionality and complexity has far exceeded our human ability for comprehension without powerful tools. While many advanced computational models and algorithms have been developed for discovering useful information in multidimensional data sets, these fully-automated approaches may not work well when users only have vague hypotheses or even no hypothesis about the data. As a result, users may have no idea on where to start the analysis process, which algorithm to choose, and how to understand the computational results. The interactive nature of data analysis processes also requires the analysis system to provide human-understandable feedback to explain the analysis results and the steps that are taken.

To loop human in the analysis process, *exploratory data analysis* resorts to visual summarization and explanations of the data to gain insight about its properties and formulate hypotheses about the data [147]. *Visual analytics* integrates automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making based on large-scalar data sets [75]. Visualization enhances human's understanding by organizing information in graphical display, offering the possibility of *visual exploration* of data for knowledge discovery and sense-making. Visual exploration strengthens human perceptual capabilities with visual interfaces to guide data navigation, actively engaging users into the exploration process to make knowledge discovery much more efficient [76].

Visual exploration of multidimensional data sets, however, remains challenging due to the heterogeneity and complexity of multidimensional characteristics. First, the multidimensional data space exceeds the limit of human comprehension, making it difficult or

even impossible to visualize the original high-dimensional data. While this problem can be alleviated by converting the data to lower dimensions through dimensionality reduction techniques [34], it is still non-trivial to navigate, compare and contrast the data items in lower dimensions, in particular when each data item contains fine-grained information to be visualized. Novel representations are thus required to display and organize data items based on the relationships of the dimensions in multidimensional data sets. Second, visual analysis of multidimensional data sets often requires investigating the hidden relationships between different dimensions and specific items to understand the multi-faceted properties of the data sets. The enormous multidimensional data space complicates the search of potentially interesting relations between dimensions and data items. Powerful and versatile visualization tools are thus needed to allow users to analyze and compare complex relations and heterogeneous structures in multidimensional data for knowledge discovery and sense-making.

1.2 Scope of Research

This dissertation focuses on investigating visualization and analysis techniques supporting multidimensional data exploration. The work of this dissertation is mainly related to three overlapping research areas (Figure 1.1): visualization (VIS), visual analytics (VA), and human-computer interaction (HCI). Here we briefly introduce the general background of the related fields to define the scope of our research. Detailed reviews of previous research efforts are provided in Chapter 2.

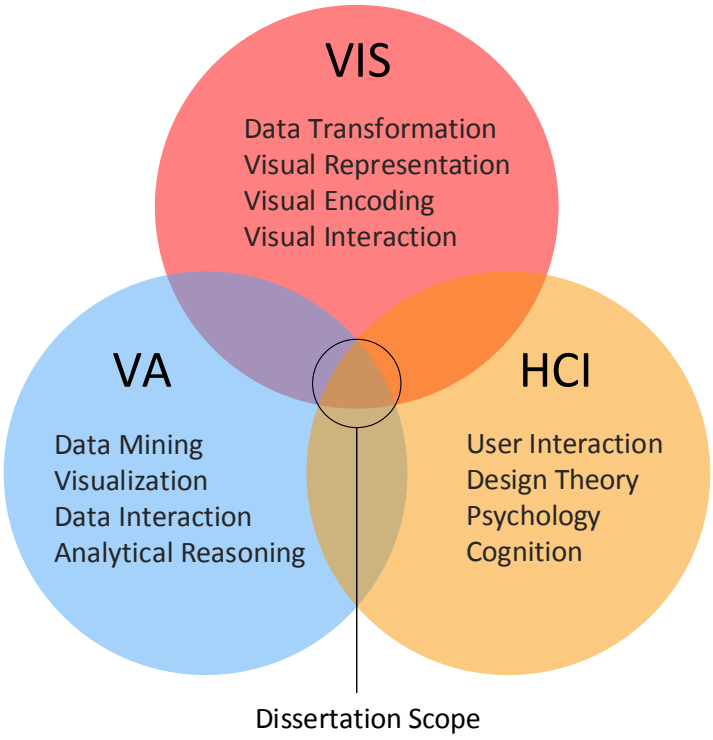


Figure 1.1: The scope of this dissertation.

Information Visualization is the research of studying the use of computer-supported, interactive visual representations of data to amply cognition [18]. Information visualization (InfoVis) and scientific visualization (SciVis) are the two major sub-areas of visualization. InfoVis studies visualizations of abstract data to reinforce human cognition through novel graphical representations and interaction techniques. Examples for abstract data are text documents of emails or messages, a database of students in a department containing names, scores and other attributes, and social networks. Since there is no inherent mapping from abstract data to space, the use of space in a visual encoding is chosen by the

designer. Munzner [109] described a central concern in InfoVis research as "determining whether the chosen idiom is suitable for the combination of data and tasks, leading to the use of methods from human-computer interaction and design". Ware [160] outlined a number of advantages for effective information visualization: (1) providing an ability to comprehend huge amounts of data, (2) allowing the perception of emergent properties that were not anticipated, (3) enabling problems with the data itself to become immediately apparent, (4) facilitating understanding of both large-scale and small-scale features of data, and (5) assisting hypothesis formation. On the other hand, the purpose of SciVis is to graphically illustrate scientific data to enable scientists to understand, illustrate, and gain insights from their data. It offers graphical representations from the results of mathematical models, computations and simulations to assist scientists in reasoning, hypothesis building and cognition. The scientific data are often continuous in both space and time. Effective scientific visualization can reveal correlations between different quantities both in space and time, create new spatial representations for visual comprehension, and open up the possibility to view the data selectively and interactively in real time. The complete data understanding process involves quantitative visual analysis at different scales in both the spatial data space and certain derived space, which motivates the use of InfoVis and human-computer interaction techniques to solve SciVis problems. This dissertation mainly focuses on the research of information visualization for designing novel interaction visualization techniques to support multi-dimensional data exploration; it also incorporates novel InfoVis and interaction techniques to provide visual exploration frameworks for exploring multivariate scientific data sets.

Visual Analytics is known as "the science of analytical reasoning facilitated by interactive visual interfaces" [136], which bridges data mining, visualization, data interaction

for knowledge discovery and sense-making. The science of analytical reasoning offers theoretical foundations of reasoning, sense-making, cognition and perception for visual analytics research to create visually enabled tools to support collaborative analytic reasoning about complex and dynamic problems. Visual representations and interaction techniques assist users to view and understand large volumes of information at once and thus facilitate knowledge discovery through visual exploration. To create such visualizations, data transformations may be used to determine the optimal way to display data, such as by creating effective 2D or 3D representations of multidimensional data [136]. Seamlessly combining visualization, human factors and data analysis techniques, visual analytics research produces practical tools that incorporate the strengths of human and machines using their respective distinct capabilities for the most effective results. Our research is related to visual analytics in developing novel visual analytical algorithms in understanding complex relationships of data items in multidimensional space.

Human-Computer Interaction is the study of how human interact with computers [19]. Researchers in HCI observe how users interact with computers, and also design new interaction techniques that enable users interact with computers in novel ways. HCI is a very broad disciplinary research that encompasses different fields such as computer science, behavioral science, cognitive psychology, design, media, art and so on. HCI research is concerned with information management, design methodology, device experiment, software system prototyping, user modeling and activity recognition. HCI lays the foundations for visualization and visual analytics by the human cognition and perception theories and frameworks. Focusing on effective visualization of multidimensional data, the research of this dissertation is guided by the theory and practice of human-computer interaction in

many aspects, and offers design guidelines and lessons derived from our studies for creating new comparative visualizations of multifaceted data sets.

1.3 Problem Statement

The fundamental goal of this dissertation is to investigate critical aspects of multidimensional data visualization and comparative analytics in assisting users in visual exploration of multidimensional data for knowledge discovery and sense-making. Specifically, we are interested in answering the following two high-level research questions:

Question 1. *How can we design multidimensional visualizations to enable effective visual summarization of multi-faceted data characteristics?* The high volume and dimensionality of data suggest that visual summarization can be invaluable to visual exploration, analysis and comprehension of large-scale multidimensional data sets. Visual summaries that highlight important aspects and characteristics of the underlying data can help users gain an effective overview of the overwhelming information embedded in multidimensional space.

Question 2. *How can we enhance the power of multidimensional visualizations with comparative analytics to allow visual identification and explanation of complex data relationships?* The relationships between different data dimensions and data items, such as substructures and correlations over various dimensions, the similarities and differences over data items and clusters, can be complex and hard to represent. Augmenting multidimensional visualization with comparative functionalities can assist users in discovering and analyzing the usually hidden multi-faceted relational patterns in multidimensional data domain.

1.4 Contributions

The main contributions of this dissertation fall into four categories: (I) novel comparative visualization techniques for multidimensional data, (II) novel multidimensional data analysis methods, (III) design implications and guidelines, and (IV) generalizable visual exploration frameworks.

First, a number of comparative visualization techniques are proposed for visualizing multidimensional data in this dissertation (shown in Figure 1.2), including:

(a) a spatially coherent comparative visualization named CorrelatedMultiples that embeds visual objects so that their distances reflect their dissimilarities;

(b) a compact visualization named TileMatrix that juxtaposes a large number of adjacency matrices for visualizing multi-faceted, time-varying networks;

(c) a composite visualization named AssociationGraph that visualizes confident scalar-level associations in multiple variables and their derived attributes;

(d) a multi-faceted visualization named BrandWheel that conveys an integrated sense of multidimensional brand personality with visual evidence and related details;

(e) a comparative visualization named BrandSediments that enables a contextual understanding of the marketplace of many brands.

Second, we contribute the following multidimensional data analysis methods:

(a) an optimization algorithm named Constrained Multi-Dimensional Scaling (CMDS) that produces an optimal proximity graph layout within a given display space;

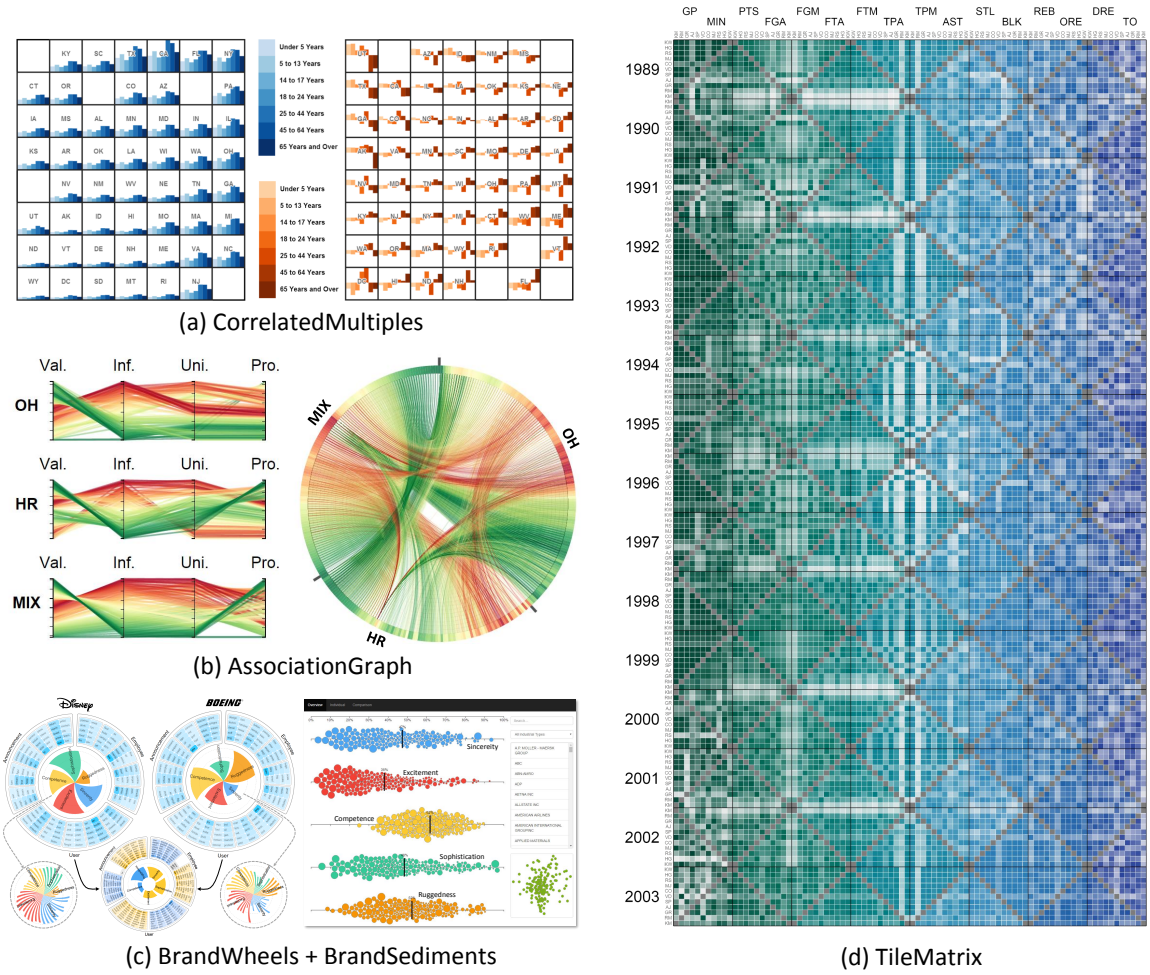


Figure 1.2: An overview of the multidimensional and comparative visualization techniques proposed in this dissertation.

(b) an association analysis method that models directional interactions between values of different variables as information flows based on association rules, and computes the informativeness and uniqueness of values based on information flows;

(c) a computational approach that is designed to assess brand personality from three driving factors of user imagery, employee imagery and official announcement on

social media, and construct a multi-faceted evidence data explaining the association between brand personality and its driving factors.

Third, the design implications and guidelines that we derived from a series of graphical perception studies and application-specific case studies consist of:

(a) design implications for choosing representations and juxtapositions of adjacency matrix visualizations;

(b) guidelines for visual exploration of scalar-level associations of different variables in the multivariate data space;

(c) design implications for developing visual analytic tools in marketing industry.

Finally, the research prototypes developed in this dissertation contribute visualization frameworks for interactive exploration of multidimensional data:

(a) a visual analytic framework with multiple interactive views for data scientists to explore the scalar values of interest with confident associations in the multivariate spatial domain;

(b) a visual analytic tool named SocialBrands for brand managers to assess and analyze public perceptions of brands on social media.

1.5 Outline

The rest of this dissertation is organized as follows.

- **Chapter 2: Background.** This chapter reviews important theories to build the foundation of this dissertation; it also surveys previous visualization techniques and analysis methods for multidimensional data and comparative analytics.

- **Chapter 3: CorrelatedMultiples: Spatially Coherent Small Multiples with Constrained Multidimensional Scaling.** This chapter introduces CorrelatedMultiples, a spatially coherent visualization based on small multiples, where the items are placed so that the distances reflect their dissimilarities. We propose a constrained multidimensional scaling (CMDS) solver that preserves spatial proximity while forcing the items to remain within a fixed region. We evaluate the effectiveness of our approach by comparing CMDS with other competing methods through a controlled user study and a quantitative study, and demonstrate the usefulness of CorrelatedMultiples for visual search and comparison in three real-world case studies.

- **Chapter 4: The Effects of Representation and Juxtaposition on Graphical Perception of Matrix Visualization.** This chapter contains two graphical perception studies to research the effects of various representations and juxtaposition designs for visualizing adjacency matrices. We investigate the effects of using square matrices and triangular matrices on the speed and accuracy of performing graphical-perception tasks. Based on human symmetric perception, we propose two alternative juxtaposition designs to the conventional side-by-side juxtaposition, and study how users perform visual search and comparison tasks regarding different juxtaposition types. With the design guidelines derived from our studies, we present a compact visualization termed TileMatrix for juxtaposing a large number of matrices, and demonstrate its effectiveness in analyzing multi-faceted, time-varying networks using real-world data.

- **Chapter 5: Association Analysis for Visual Exploration of Multivariate Scientific Data Sets.** This chapter describes a novel association analysis method that

guides visual exploration of scalar-level associations in the multivariate context. We introduce the concepts of informativeness and uniqueness to describe how information flows between scalars of different variables and how they are associated with each other in the multivariate domain. Based on scalar-level associations represented by a probabilistic association graph, we propose the Multi-Scalar Informativeness-Uniqueness (MSIU) algorithm to evaluate the informativeness and uniqueness of scalars. We present an exploration framework with multiple interactive views to explore the scalars of interest with confident associations in the multivariate spatial domain, and provide guidelines for visual exploration using our framework.

- **Chapter 6: SocialBrands: Visual Analysis of Public Perceptions of Brands on Social Media.** This chapter presents a design study of a marketing analytical tool for brand managers to understand public perceptions of brands on social media. The proposed system, named SocialBrands, leverages brand personality framework in marketing literature and social computing approaches to automatically compute the personality of brands from three driving factors (user imagery, employee imagery, and official announcement) on social media, and construct an evidence network explaining the association between brand personality and driving factors. These computational results are then integrated with new interactive visualizations to help brand managers understand the personality traits and the associated social media factors. We demonstrate the usefulness and effectiveness of SocialBrands through a series of user studies with brand managers in an enterprise context. Design lessons are also derived from our studies.

- **Chapter 7: Conclusions.** This chapter summarizes the contributions of this dissertation, describes implementation details of the research prototypes, and outlines several promising directions for future work.

Chapter 2: Background

This chapter introduces the theoretical background of visualization, and reviews the literature of visualization research with the focus on the following areas: multidimensional data visualization, comparative visualization, and interactive visual exploration. Additional specific and more detailed related works will be discussed in the corresponding chapters (Chapters 3-6).

2.1 Theory and Foundation

Visualization has been an effective way to communicate both abstract and concrete information for centuries. In the book "The Visual Display of Quantitative Information" [146], Edward Tufte described excellent visualization as:

(a) well-designed presentation of interesting data;

(b) consisting of complex ideas communicated with clarity, precision and efficiency;

(c) nearly always multivariate;

(d) telling the truth about the data;

(e) giving the greatest number of ideas in the shortest time with the least ink in the smallest space.

Historic examples of excellent visualization include the Ptolemy's world map [137] (Figure 2.1) and the map of Napoleon's invasion [106] in Russia (Figure 2.2). Developments in technologies such as printing, reproduction and computing have expanded the use of graphics to advance visualization. The rise of visual thinking and human-computer interaction also contributes to the development of visualization and visual analytics. Visualization today has ever-growing applications in science, engineering, education, medicine, finance, and so on. In this section, we review the main theories of visualization and visual analytics to build the foundations of this dissertation.

Figure 2.1: The Ptolemy's world map [137], built from Ptolemy's Geographia, indicating the countries of "Serica" and "Sinae" (China) on the right, beyond the island of "Taprobane" (Sri Lanka) and the "Aurea Chersonesus" (Southeast Asian peninsula).

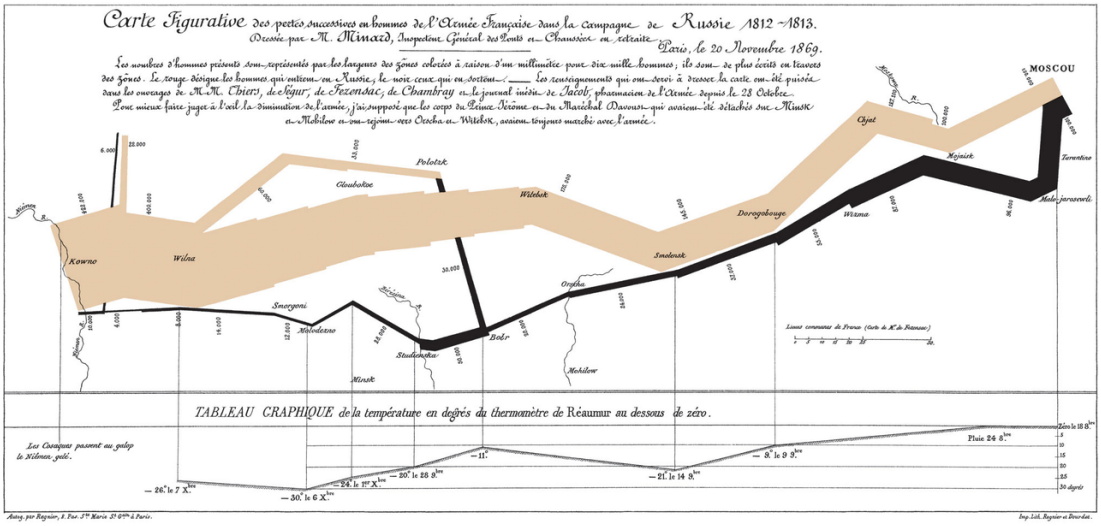


Figure 2.2: The map of Napoleon's invasion by Charles Joseph Minard [106], showing the decreasing size of the Grande army as it marches to Moscow (brown line, from left to right) and back (black line, from right to left) with the width of the line encoding the size of the army.

2.1.1 Graphical Perception

The effectiveness of visualization primarily depends on how it is perceived by viewers. Edward Tufte [146] emphasized that effective visualization should be self-explanatory: "the translation of visual to verbal is quickly learned, automatic, and implicit — so that the visual image flows right through the verbal decoder initially necessary to understand the graphic". Graphical perception is defined as the visual decoding of information encoded on graphical displays [21]. The graphical dimensions for visual encoding are known as *visual variables* or *visual channels*. Visual variables were first classified by Jacques Bertin [13] as position, size, shape, color brightness, color hue, orientation, and grain. Morrison [107] suggested the addition of color saturation and arrangement, and MacEachren [97, 98] proposed three more: fuzziness, resolution, and transparency. The list can be further expanded

by including some of the visual primitives examined by Cleveland and McGill in their seminal work on graphical perception [21], including angle, volume, and curvature. Carpendale [20] further considered motion, depth and occlusion as visual variables that go beyond conventional static and 2D display. Bertin [13] and Carpendale [20] offered detailed analysis and usage suggestions of the visual variables based on the ranks of their effectiveness. Munzner [109] classified the visual variables into *magnitude variables* (for encoding ordered or numerical attributes) and *identity variables* (for encoding categorical attributes), and characterized the effectiveness of visual variables (in Figure 2.3 and Figure 2.4) by the following criteria:

- **Accuracy**: how close is human perceptual judgement to objective measurement?
- **Discriminability**: whether the differences between objects are perceived as intended?
- **Separability**: how much the potential interactions are between the visual variables?
- **Visual popout**: how easily a distinct object can stand out from many others?
- **Perceptual grouping**: how strong the perceptual grouping cue is of a visual channel?

A visualization may contain many groups of visual objects, with the objects themselves consisting of smaller parts. *Gestalt principles*, proposed by Max Werteimer [78], describe how human tend to perceive visual elements as groups (gestalt means 'shape' or 'form' in German). In visual perception, such groups are the regions of the visual field whose portions are perceived as joined together, which are segregated from the rest of the visual field (as shown in Figure 2.5). The Gestalt principles are described by the following laws:

- **Proximity**: objects that are near each other tend to be perceived as groups;
- **Similarity**: similar objects tend to be integrated into groups;

- **Continuity**: objects that are aligned with each other tend to be integrated into groups;
- **Closure**: objects tend to be grouped together if they are parts of a closed figure;
- **Symmetry**: symmetrical objects tend to be grouped to form a coherent shape;
- **Common fate**: objects moving together tend to be perceived as groups;
- **Past experience**: objects tend to be grouped together if they were often together;
- **Good gestalt**: objects tend to be perceptually grouped together if they form a regular, simple and orderly pattern.

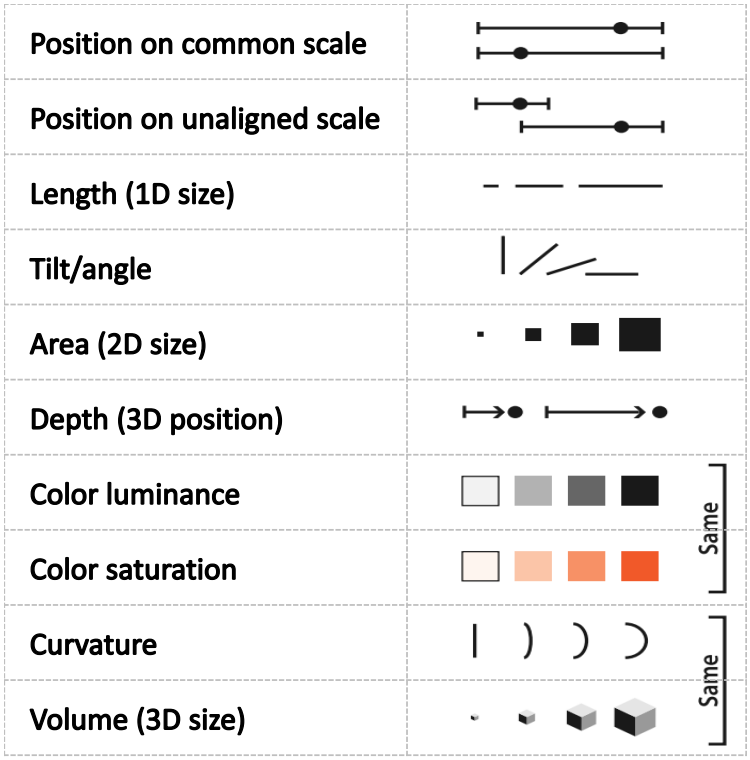


Figure 2.3: Ranking effectiveness of the visual variables of ordered attributes (decreasing effectiveness from top to bottom). Adapted from Figure 5.6 in [109].

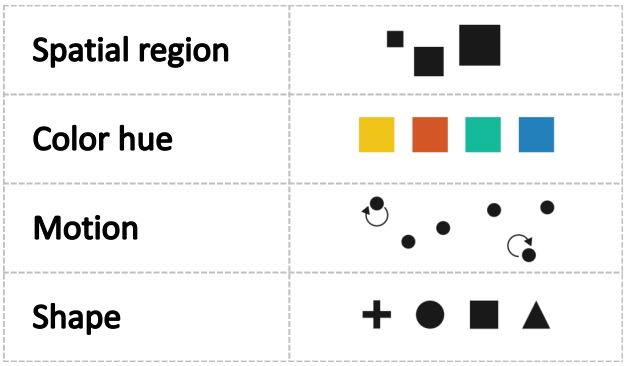


Figure 2.4: Ranking effectiveness of the visual variables of categorical attributes (decreasing effectiveness from top to bottom). Adapted from Figure 5.6 in [109].

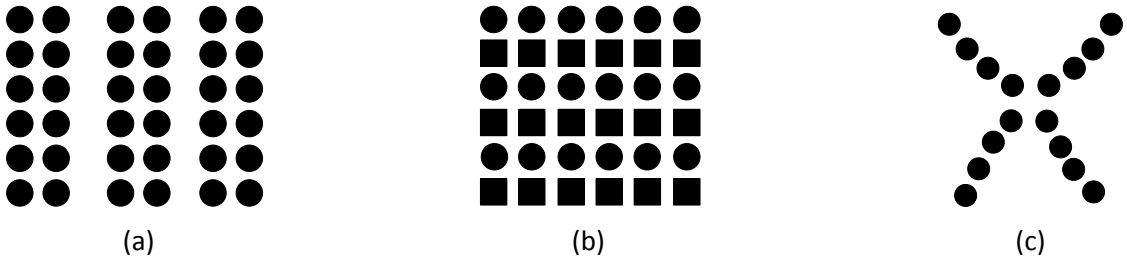


Figure 2.5: Illustration of Gestalt principles. (a) The law of proximity: we see three columns as the lines of circles near each other appear to be grouped together. (b) The law of similarity: we see lines of circles and lines of squares grouped together. (c) The law of continuity: we see smooth and continuous lines of circles.

2.1.2 Visualization Design

Visualization design is often a challenging task because there exist many possibilities in the design space and most will be ineffective for a specific usage context [109]. A confusing design can be poorly perceived by the viewers. A possible design that is perceivable in a certain setting may be a poor match regarding an intended analysis task. Only a very small number of design possibilities are reasonable candidates, and of those only an even smaller portion are effective designs. As illustrated in Figure 2.6, a good visualization design

should start from a large *known space* (the known possible design choices) and a large *consideration space* (the considered design choices). When the known and consideration spaces are small, the risk of choosing a poor solution is high. A fundamental principle of visualization design is to propose multiple design candidates (forming a possibly large *proposal space*) and select the best one, instead of immediately focusing on one solution without considering any alternatives.

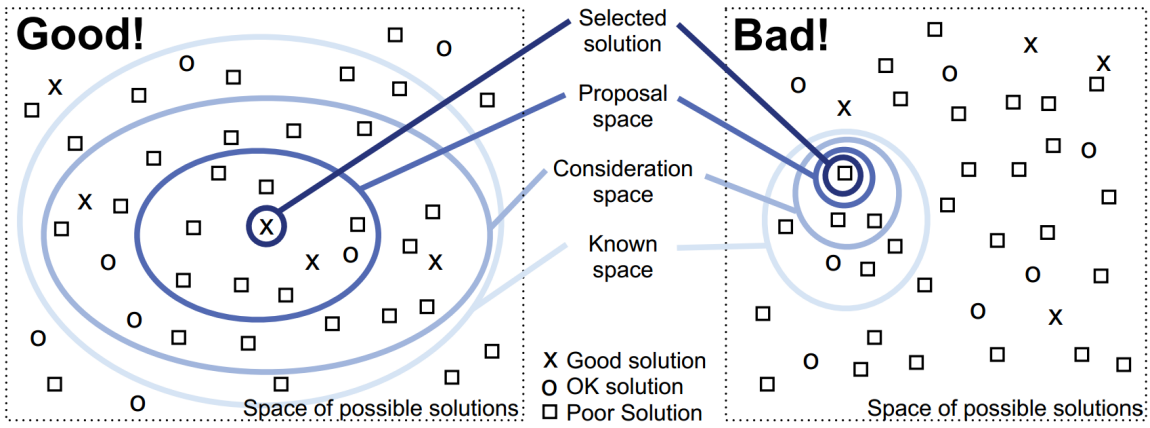


Figure 2.6: A search space metaphor for visualization design. After Figure 1.5 in [109].

.

The best design cannot be easily determined by a simple optimization process as there exist trade-offs between design alternatives. The characterization of trade-offs between design alternatives has stimulated many research efforts in visualization research. Edward Tufte [143, 144, 145, 146] summarized a series of design principles on creating effective visualization, such as focusing on the data rather than data-containers, providing more relevant information within the eye-span, minimizing visual clutter and confusing encodings, and maximizing the *data-ink ratio* (the ink for encoding the data scaled by the total ink

used to print the graphic) with reasons. Munzner [109] further provided general principles to guide the use of visual variables in visualization design. The *expressiveness principle* suggests that the visual encoding should express all of and only the information in the data attributes: ordered data should be sensed as ordered; unordered data should not imply an misleading ordering that does not exist. The *effectiveness principle* suggests that the importance of the data attribute should match the saliency of the visual variable: the most important attributes should be encoded with the most effective visual variables, while the decreasingly important attributes can be matched with decreasingly effective variables. Colin Ware [159] suggests to encoding multi-dimensional information using separable visual variables such as color, shape and motion (an illustrative example is shown in Figure 2.7).

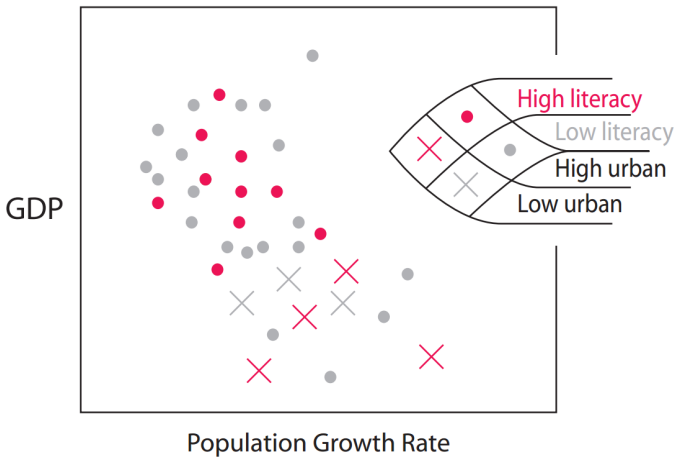


Figure 2.7: An illustrative example of encoding multi-dimensional information using separable visual variables in a scatterplot design. Two visual variables — color and shape — are used to categorize the degrees of literacy and urbanity of the data points. From [159].

In this dissertation, we base our design process on these principles to justify the choices in our visualization designs.

2.1.3 Visualization Model

To characterize the computational process of creating a visualization, Card et al. [18] presented the *visualization pipeline* that transforms information into a visual representation (Figure 2.8). To start with, the input raw data is first transformed into a well-organized data format, which typically consists of a set of data items with associated data attribute values. Various data processing approaches, such as data mining or machine learning techniques, can be applied to gain insights from the input data and derive new data as needed. After that, a critical step of the visualization process is to map the processed data into a certain visual representation, which contains visual metaphors that are designed to represent the data items. The next step transforms visual representations into views, which display the visual representations on the display screen and provide various view transformations for navigation. The views are presented to users through the human visual system, which enables users to perceive the view to reconstruct the underlying information. In the meanwhile, users can interact with any of the steps in the visualization pipeline to change the resulting visualization, and thus make further interpretations.

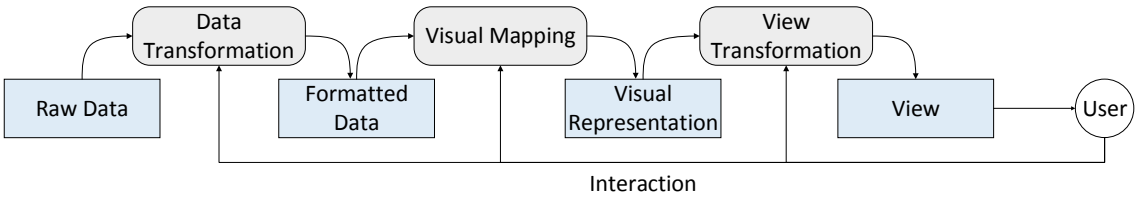


Figure 2.8: Visualization Pipeline, adapted from [18].

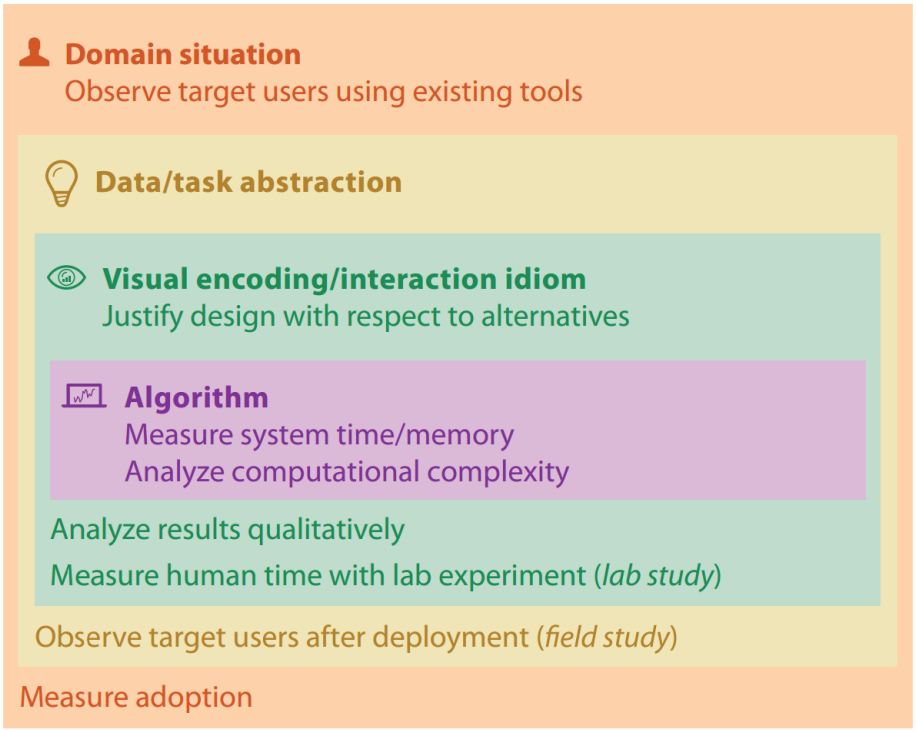


Figure 2.9: The nested model for visualization design and validation consists of four levels: domain situation, data/task abstraction, visual encoding/interaction idiom, and algorithm. After Figure 4.1 from [109].

To guide visualization design and validation, Munzner [109] proposed a nested model that decomposes the visualization design process into four nested levels (Figure 2.9): (1) domain situation — observing target users using existing tools and identifying problems; (2) task and data abstraction — mapping domain-specific problems and data into forms that are independent of the domain; (3) visual encoding and interaction idiom — specifying and justifying design choices with respect to alternatives; and (4) algorithm — implementing the visualization design computationally. The nested model can also be used to validate a possible visualization design from the inner levels to the outer levels: (1) measuring system time and memory, and analyzing computational complexity; (2) analyzing results

qualitatively; (3) measuring human time with lab experiment; and (4) observing target users after deployment.

2.1.4 Visual Analytics

Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces [136]. According to the visual analytics research and development agenda "Illuminating the Path" by Thomas and Cook [136], visual analytics is a multidisciplinary field that consists of the following aspects: (1) analytical reasoning techniques — gaining deep insights from massive data that directly support assessment, planning, and decision making; (2) visual representations and interaction techniques — utilizing human eye's broad bandwidth pathway into the mind to enable users to see, explore and understand large amounts of information; (3) data representations and transformations — transforming data in ways that support visualization and analysis; and (4) production, presentation and dissemination techniques — communicating the analysis results in the appropriate context to a variety of audiences.

A visual analytics approach can be understood as a way to integrate automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making based on large-scalar data sets [75]. On the one hand, automatic analysis methods are part of the *knowledge discovery and data mining* discipline with strong theoretical foundations. However, the analysis results can be difficult for domain users to understand because the algorithms do not take into account relevant expert knowledge. On the other hand, while visualization approaches can produce acceptable results for small data sets, they may fail when the given data is too large to be handled by a domain user. The fundamental goal of visual analytics is to keep domain users in the analysis loop, with

the ability to interact and steer the analysis process. The design of visual analytics tools is desirable to root in a theoretical foundation, amplify human's understanding of the domain tasks with interactive visualization, and loop human's knowledge into the theoretical framework [135].

Keim et al. [75] provided a visual analytics model of sense-making loop (shown in Figure 2.10), following a visualization schema developed by van Wijk [153]. Initially, the raw data is processed through statistical and mathematical transformations. The analytical process then enters a sense-making loop where the user can gain knowledge by discovering new insights from the perceived visualization images. In the meanwhile, the user can assess and steer the visualization itself, develop new hypotheses, adjust analytical and exploratory techniques, and interact with the visual representation for achieving more focused and more appropriate visualization.

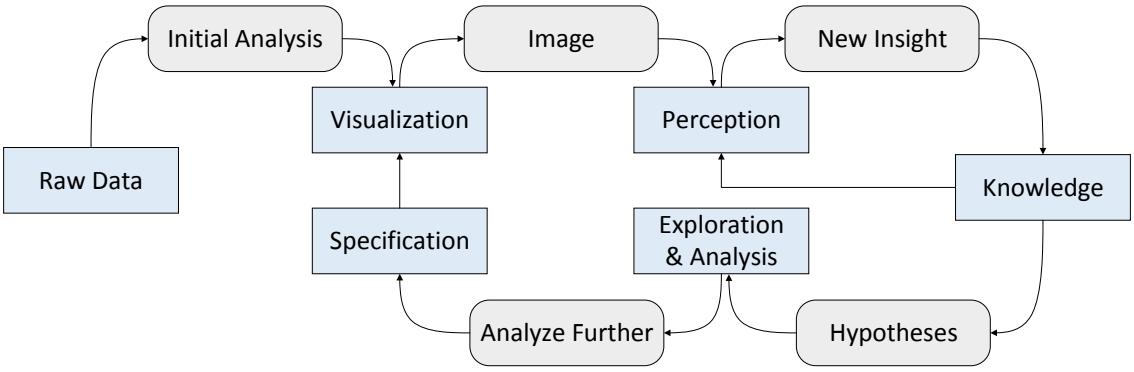


Figure 2.10: The visual analytics model of sense-making loop, adapted from [75].

2.2 Multidimensional Data Visualization

In the book "Envisioning Information" [144], Edward Tufte noted that "all the interesting worlds that we seek to understand are inevitably and happily multivariate in nature". Multi-dimensional data visualization is a fundamental research topic in both scientific and information visualization. Scientific simulations often create multiple variables describing different physical properties within the same spatial domain. Usually, certain scalar values from a subset of variables carry a greater importance to the understanding of the underlying phenomena than others [73]. The integration of abstract data from multiple facets is also common in information visualization when visualizing relational databases with may attributes and categories. Wong and Bergeron [162], Fuchs and Hauser [36], and Kehrer and Hauser [73] gave extensive surveys of multivariate data analysis and visualization. In this section, we briefly review the major visual representations and analysis methods for multivariate data visualization.

2.2.1 Representation of Multidimensional Data

A scatterplot is a basic visual representation to display values for typically two variables using Cartesian coordinates. Built upon on scatterplot, a scatterplot matrix [53] is created by juxtaposing dimensions vertically and horizontally, and then plotting a scatterplot for each pair of dimensions (Figure 2.11). While scatterplot matrix is a popular multidimensional visualization for summarizing relationships of dimensions, it is mostly effective for encoding pairwise correlations and the more complex multidimensional correlations can be difficult to interpret. Parallel coordinates plot (PCP) [61] displays the multidimensional data points as polylines spanning a set of parallel axes, each representing one dimension (Figure 2.12). Star Coordinates [71] uses a circular layout of dimension

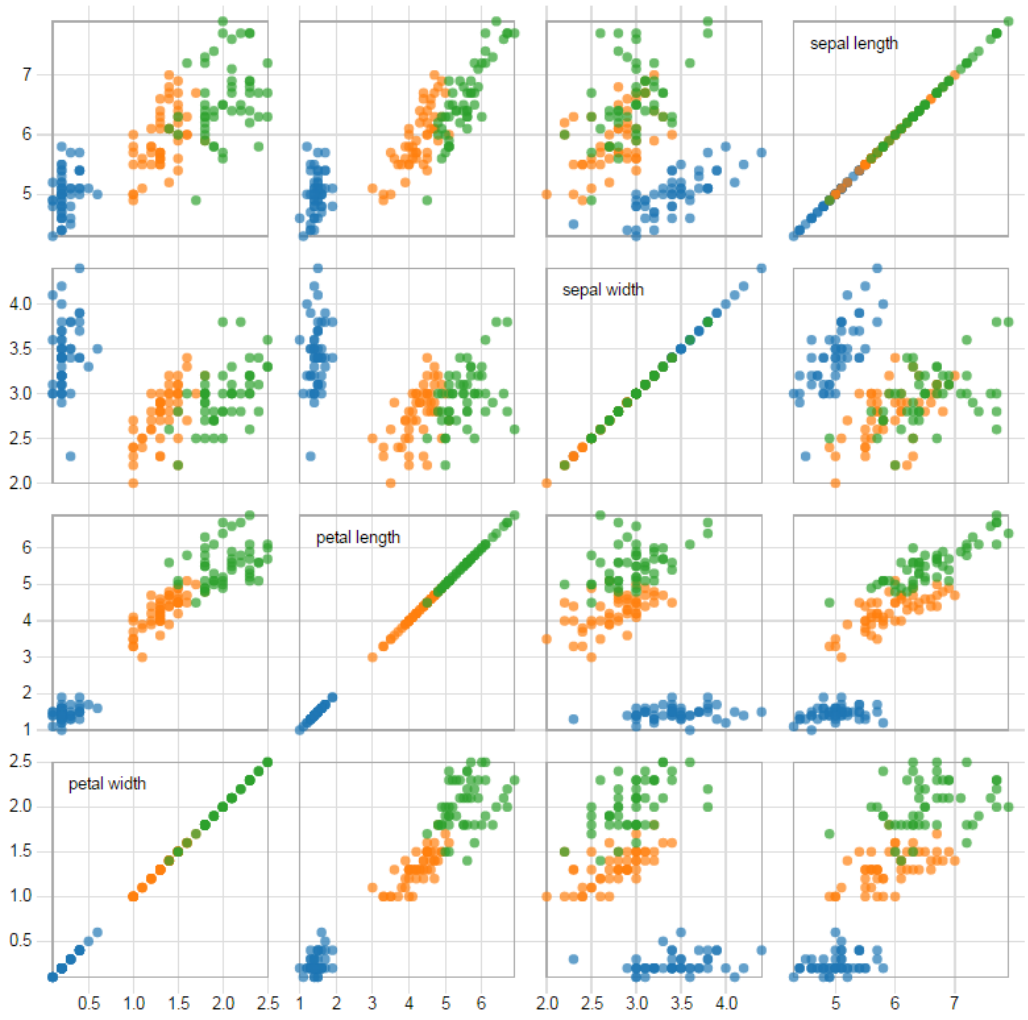


Figure 2.11: An illustration of a scatterplot matrix using the Iris data set [33], containing four dimensions: petal width, petal length, sepal width, and sepal length.

axes to project high-dimensional data points onto a 2D plane based on linear combinations of a set of low-dimensional vectors. Similarly, RadViz [58] models the dimensions as a physical spring system across the dimension axes to generate nonlinear mappings of high-dimensional data points onto a 2D plane. While these visualization techniques are intuitive

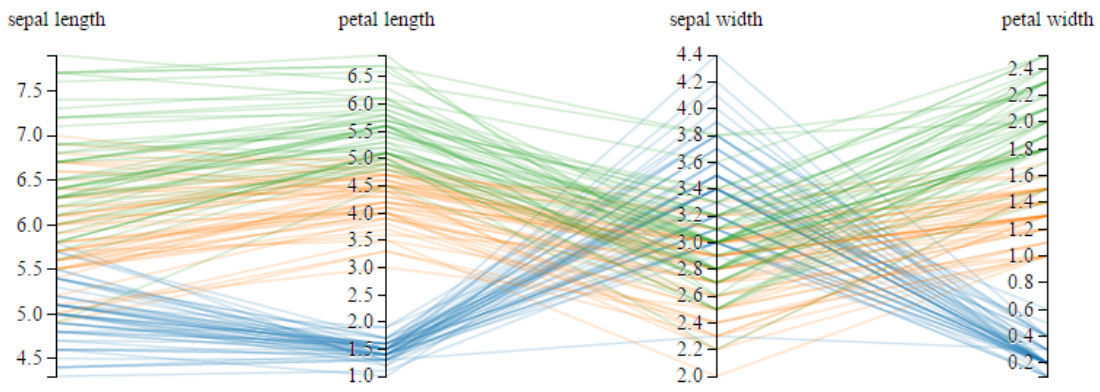


Figure 2.12: An illustration of a parallel coordinates plot using the Iris data set [33], containing four dimensions: petal width, petal length, sepal width, and sepal length.

to explore the multidimensional data relationships, most are not scalable to data with very high dimensionality.

To visualize multidimensional data that is large in dimensionality, dimensionality reduction techniques, which define procedures that map numerical multidimensional data onto a low-dimensional display, are often used to represent multidimensional data graphically. Dimensionality reduction methods can be categorized as linear or nonlinear. Linear techniques, such as principal component analysis (PCA) [67], statistical biplots [37] and projection pursuit [68], map high-dimensional data points onto a low-dimensional space by a simple matrix-vector product. Nonlinear techniques construct sophisticated mappings to embed high-dimensional data points on a low-dimensional manifold as faithfully as possible, which account for the relationships such as similarities and differences between data points. Popular methods include multidimensional scaling (MDS) [82, 83, 84], self-organizing map (SOM) [80] and t-distributed stochastic neighborhood embedding (t-SNE) [152]. As shown in Figure 2.13, while a low-dimensional embedding produced by

dimensionality reduction is easier to visualize, the original dimensionality information is lost, and the intrinsic relationships between dimensions are not explicitly available.

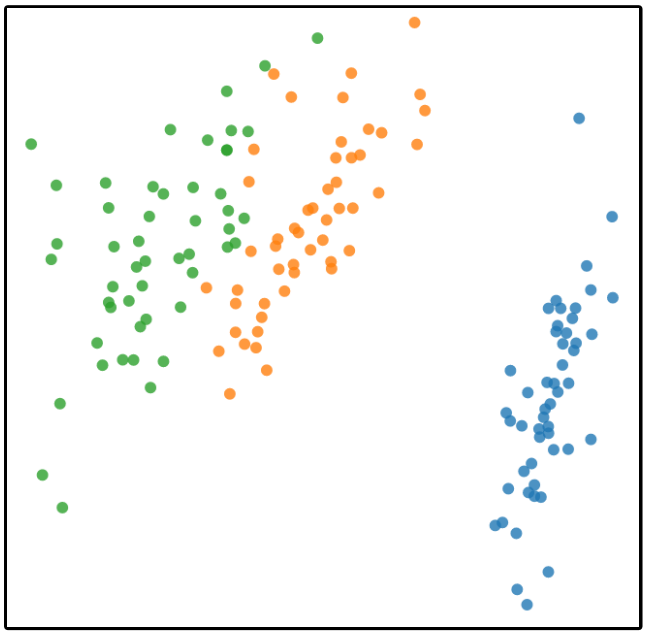


Figure 2.13: An illustration of the 2D embedding of the Iris data set [33] using multidimensional scaling (MDS).

Multiple representations can be integrated or linked to show the multi-faceted characteristics of multidimensional data. Many research efforts have been made in developing hybrid visualization techniques for multidimensional data. Schmid and Hinterberger [123] combined scatterplot matrices, parallel coordinates plots, permutation matrices and curve display together. Wong and Bergeron [163] linked parallel coordinates plots with scatterplots matrices. Yuan et al. [170] integrated parallel coordinates with scatterplots, using MDS to convert multiple axes into a single 2D plot. Turkay et al. [148] and Yuan et al. [171]

linked the data items plot and the dimensions plot together to support simultaneous exploration of data correlations and dimension correlations in multidimensional data.

2.2.2 Analysis of Multidimensional Data

A fundamental research problem in analysis of multidimensional data is to study the relationships of dimensions (variables or attributes) in the multidimensional data space. Sauber et al. [122] introduced local correlation coefficients to visualize relationships of variables in multidimensional data sets. Gosink et al. [47] took normalized dot product between two gradient fields from two variables to derive correlation fields, which were then used to analyze the variable interactions with a third variable. Janicke et al. [64] adapted local statistical complexity from finite state cellular automata to identify informative regions in multidimensional data. Nagaraj et al. [110] proposed a gradient-based measure that reveals the relationships of variables for the purpose of comparative visualization. Wang et al. [156] employed transfer entropy to study the causal relationships between variables. Tatu et al. [133] developed an interactive visualization system for exploring interesting subspaces of high-dimensional data. Watanabe et al. [161] applied biclustering techniques to extract feature subspaces from multidimensional data. Biswas et al. [14] employed the surprise and predictability metrics to measure the specific mutual information between a scalar value and a variable.

While most existing methods offer insights into the average relationships between variables, little study has focused on the specific relationships between different scalar values in different variables. In this dissertation, we present a bottom-up approach to analyze scalar-level relationships in different variables for identifying the informative and unique scalars in multivariate data sets. Our approach copes with more than two variables by using

two new metrics, informativeness and uniqueness, to reflect two-way interactions between one scalar value and other scalars of different variables.

2.3 Comparative Visualization

Visual analysis often involves visual comparison of multiple objects. Comparison tasks appear across many domains such as finance, sociology, biology, climatology, and network analysis for various types of data including texts, graphs, tabular data and volumetric data. Introduced by Pagendarm and Post [112], comparative visualization plays an essential role in comparative data analysis, which focuses on analyzing relationships (such as similarities and differences) between data dimensions, data instances, data sources, or data at different time steps. Gleicher et al. [42] described juxtaposition, superimposition, and explicit encoding as three generic design approaches to comparative visualization, and suggested that multiple designs can be combined to created hybrid comparative visualization. One major goal of this thesis is to provide a thorough understanding of how to design comparative visualization (particularly juxtaposed visualization) for visual exploration of multidimensional data. In this section, we review the three basic designs of comparative visualization, with a focus on juxtaposition. From now on, we denote the basic visualization of one object as a *visual object*, while a specific comparative visualization is formed by compositing multiple visual objects.

2.3.1 Juxtaposition

Juxtaposition is an effective visual design that encourages side-by-side comparison of multiple visual objects with little additional visual clutter. A juxtaposed visualization predominantly relies on the use of the viewer's memory and attention shifts to make connections between repeated visual objects [42, 65]. Jacques Bertin [13] and Edward Tufte [143]

introduced *small multiples*, the most popular juxtaposition design that arranges a set of visual objects in a grid layout to encourage comparison (as shown in Figure 2.14). Edward Tufte [146] considered small multiples to be *economical* for comparative analysis: "once viewers understand the design of one slice, they have immediate access to the data in all the other slices; the constancy of the design allows the viewer to focus on changes in the data rather than on changes in graphical design". The effectiveness of small multiples has been extensively studied in the literature of information visualization. Heer et al. [54] measured the effects of chart size and layering on user performance when performing visual discrimination and estimation tasks. Fuchs et al. [35] investigated user performance of different temporal glyph designs in a small multiple setting. Archambault et al. [5] found that small multiples gave significantly faster performance than animation for understanding dynamic graphs. Robertson et al. [118] also showed that small multiples are more accurate and effective than animation for trend analysis.

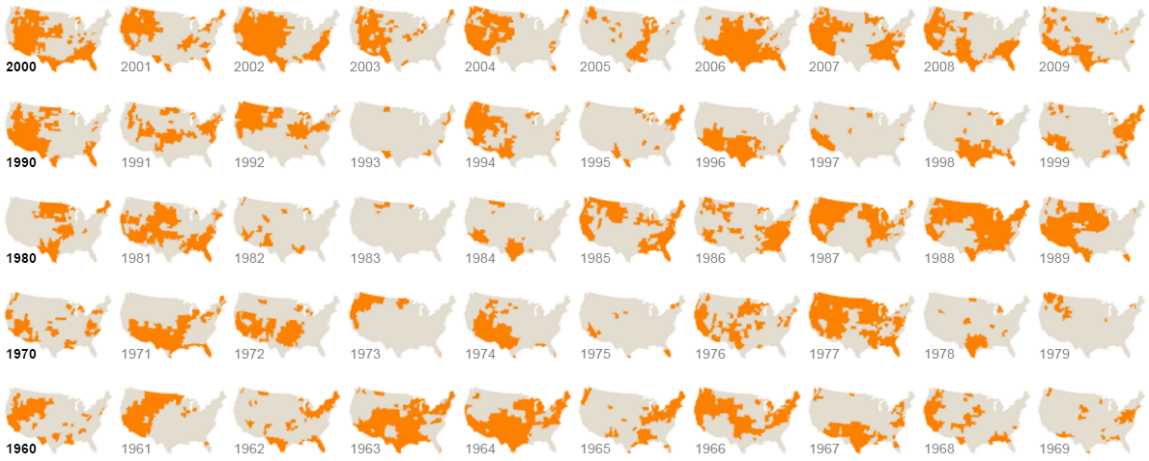


Figure 2.14: An illustration of multiple juxtaposed U.S. maps, highlighting dry areas in June from 1970 to 2009. Adapt from [139].

When too much of the comparative burden is added onto the viewer's mental effort, he or she can fail to detect changes even if the information is well represented in the visualization [125]. Therefore, the order and layout of small multiples are critical to the effectiveness of juxtaposed visualization. Edward Tufte [144] suggested that "if the visual task is contrast, comparison, and choice — as so often it is – then the more relevant information within eye span, the better". As early as 1983, Jacques Bertin [13] considered the possibility of reordering the items to highlight interesting relationships. Since then, organizing objects based on their similarity has been an important research topic for creating effective juxtaposition. Gomez-Nieto et al. [46] showed that 2D similarity-based visualization allows users to identify related items faster than using 1D ranked list. Haroz and Whitney [51] also found that 2D grouping significantly improves task performance of visual search. However, grouping similar items while efficiently utilizing the display space is non-trivial. Self-organizing maps (SOM) [79] projects high-dimensional data to a lower dimension while preserving relationships in the input data [79]. Unfortunately, multiple items may be projected onto the same position as the best-matching unit on the low dimensional map, which can result in large overlaps [129]. Itoh et al. [62] proposed a hybrid space-filling and force-directed layout to visualize multiple-category graphs, but some space between groups is wasted. IncBoard [115] incrementally updates a 2D grid to place similar items together, which also suffers from inefficient space utilization. Kehrer et al. [74] exploited the hierarchical structure of small multiples to support multi-category comparison. While many research efforts have been made to generate spatially coherent layouts when removing overlaps [24, 29, 38, 46, 59, 128], they did not consider the boundary of the display space as an essential constraint. Although the items after applying these

overlap removal methods can be uniformly scaled down to fit the display space, it can lead to readability issue and inefficient space utilization[94].

To overcome this problem, space-filling techniques offer an appropriate option to maximize the utilization of display space while creating juxtaposition visualization. Both greedy approaches [119, 129, 130, 164] and optimization methods [32, 45, 151] have been proposed to generate spatially coherent grid layouts. Rodden et al. [119] searched a locally optimal grid position based on spiral heuristics. Wood and Jason [164] ordered grid cells with two-dimensional consistency. Strong and Gong [129] proposed a multi-dimensional pseudo-sorting algorithm to minimize local dissimilarity. Strong et al. [130] further proposed an algorithm to maximize the similarity-proximity correlation. Eppstein et al. [32] showed that greedy approaches such as [164] can result in a relatively large displacement and thus considered less optimal. However, optimal matching is computationally expensive. MIOLA [45] used mixed integer optimization to ensure a grid-like layout, which has exponential worst case complexity. The optimization algorithm proposed in this dissertation can produce grid layouts comparable in quality to the best known algorithm [32], but more efficiently.

2.3.2 Superimposition

In contrast to position, superimposition overlays many visual objects in a single visualization. The visual objects share the same display space, which facilitates direct visual comparison of similarities and differences. Superposition is very popular in chart-style visualizations in the form of overlaying one chart on top of another. For instance, multiple line charts can share the display space in a single visualization, as shown in Figure 2.15. When visual objects are aligned in such a shared display space, it is easier to perceive subtle

differences across objects due to the prominent effectiveness of the visual channel *aligned spatial position* [109]. Javed et al. [66] evaluated user performance for visual comparison, slope, and discrimination tasks for multiple line graph visualizations, and showed that superimposition is typically more efficient for comparisons over a smaller visual span while juxtaposition is generally more efficient for comparison with a large visual span.

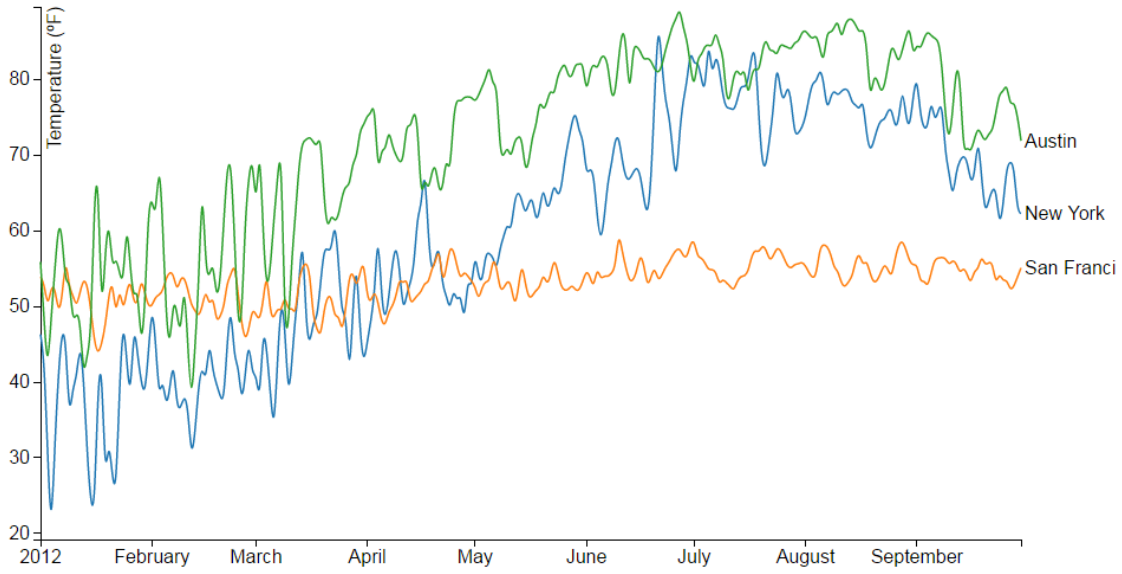


Figure 2.15: An illustration of superimposed line charts, showing the daily average temperatures of New York, San Francisco and Austin, adapted from [16].

When the superimposed visual objects are complex and dense (such as networks and images), or when the number of overlaying objects is large, visual clutter and occlusion can hinder the perception of similarity and differences among objects. Baldassi et al. [7] showed that visual clutter can mislead users to problematic judgments and more erroneous decisions. A systematic analysis of clutter reduction techniques in information visualization can be found in a taxonomy proposed by Ellis and Dix [31]. To reduce clutter in

superimposition design, Hagh-Shenas et al. [49] studied the effects of color blending (a linear combination of individual colors) and color weaving (individual colors are displayed side-by-side in a high frequency texture that fills the display space) for mixing colors in a single visualization; color weaving was recommended for overlaying more than two visual objects. Miller [105] introduced attribute blocks to take alternating samples from different visual objects to achieve a superimposition effect. Malik et al. [101] exploited a hexagonal decomposition of the display space for superimposing a relatively larger number of images at different scales.

Juxtaposition and superimposition can be combined to design hybrid comparative visualizations. Javed and Elmqvist [65] identified overloading and nesting as two such hybrid visualization designs in addition to juxtaposition and superimposition: the overloading design utilizes the space of one visual object to display another for achieving a compact visualization; the nesting design embeds the contents of one visual object inside another for augmenting a particular visual representation with additional visual mapping. For instance, Figure 2.16 nests multiple migration categories in a geographical visualization of the U.S. states.

2.3.3 Explicit Encoding

Another alternative of comparative visualization is to compute the relationships between objects and provide explicit visual encoding of the relationships. In other words, data is transformed such that the comparative relationships of objects can be visualized explicitly. While this comparative visualization design saves the viewer's effort in making comparison judgements, it requires the knowledge of pre-defined relationships to make a certain transformation from the original data into relationships.

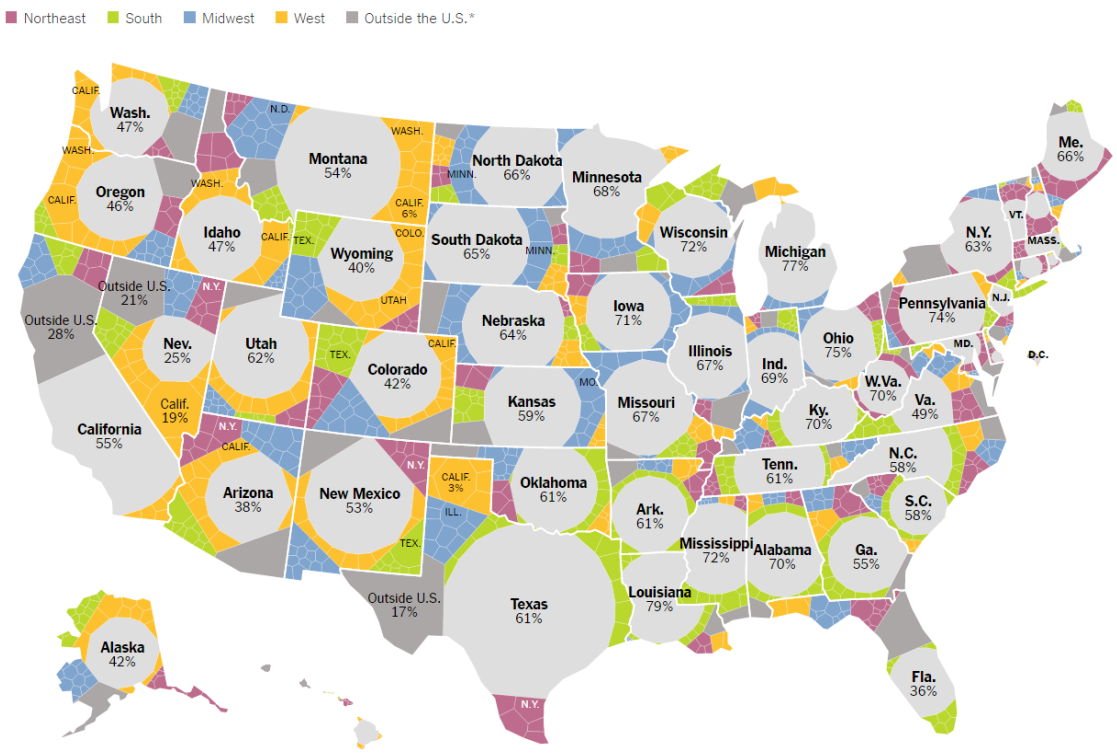


Figure 2.16: An illustration of explicit encoding of multiple migration categories in a geographical visualization of the U.S. states. After [140].

Since explicit encoding focuses on visualizing the relationships of objects, the original data objects are often removed from the resulting visualization. Loss of the context of the original objects can make it difficult for a viewer to relate the displayed relationships back to the data objects themselves. Therefore, explicit encoding is often combined with the other two comparative designs as augmented visual encoding: an explicit encoding of the data objects may be juxtaposed by a visualization of the original data objects, or superimposed on top a visualization of the original data objects. For example, Wang et al. [157] designed a nested PCP that integrates juxtaposition, superimposition and explicit encodings in a single view for comparative data visualization and analysis.

2.4 Interactive Visual Exploration

Exploratory data analysis, promoted by John Tukey [147] in statistics, encourages analysts to explore data sets through visual summarization and explanation of the main characteristics, and thus formulate hypotheses that could lead to new data collection and experiments. Interaction with visualization, which enables users to dynamically manipulate visual representations of data according to their exploration purposes, is critical to visual exploration and comparative analytics of multidimensional data. In this section, we review the fundamental guidelines and typical interaction techniques for visual analytics of multidimensional data.

The classic visualization seeking mantra "overview first, zoom and filter, detail on demand", proposed by Ben Shneiderman [124], describes a typical visual exploration process and provides the basic guidelines for designing effective visualization systems. In particular, an overview of the data should be first created to initiate the visual analysis process. After that, the user can drill down into the region of interest via zooming and filtering, which enables the user to view information details on demand. Furthermore, Ben Shneiderman [124] outlines three important visualization tasks: (1) relate — viewing relationships among items, (2) history — keeping a history of actions to support undo, replay and progressive refinement, and (3) extract — allowing extraction of sub-collections and of the query parameters. Keim et al. [75] extended this information seeking mantra to visual analytics: "analyze first, show the important, zoom, filter and analyze further, details on demand". First, analytical approaches should be seamlessly combined with advanced visualization techniques. In particular, analysis procedures should be conducted before visualization to extract important features and insights for assisting a user to proceed the exploration. Second, it suggests that a data exploration process is iterative in the sense that

users should be able to take actions after gaining new knowledge about the data to analyze further and deeper. Users' immediate feedback in return should be taken into consideration to change the visualization and analytical algorithms within the analytical workflow.

Introduced by Becker and Cleveland [10], brushing became one of the most popular interactions for exploring data of interest. Martin and Ward [103] extended brushing to high-dimensional data space. Janicke et al. [63] proposed to brush an attribute space derived from the high-dimensional data space. Coordinated and multiple views are an effective design for visual exploration through different presentations, where the changes in one representation will update the other views immediately [8]. Many examples are given by Roberts [117] in a survey of coordinated and multiple views in exploratory visualization. To represent relationships of variables, node-link diagrams and PCP are commonly used, as in Wang et al. [156], Yang et al. [167], and Biswas et al. [14]. Tominski et al. [142] proposed shine-through and folding interactions for visual comparison besides side-by-side comparison. Perin et al. [114] showed that simple interactions such as zooming and panning can substantially improve efficiency of comparative visualization of multiple line charts and horizon graphs [121]. In this dissertation, interactions such as brushing and linking are extensively used to enhance visual comparison in various comparative analysis tasks as well as knowledge discovery in different visual exploration processes.

2.5 Summary

While we described the scope of the research in the previous chapter, this chapter reviews the literature related to various aspects of this dissertation research. Inspired by many aspects in the literature reviewed in this chapter, we present specific topics in the upcoming chapters, including multidimensional visualization techniques, comparative visualization

techniques, multidimensional data analysis methods, design studies for domain-specific applications, and generalizable visual exploration frameworks.

Chapter 3: CorrelatedMultiples: Spatially Coherent Small Multiples with Constrained Multidimensional Scaling

3.1 Motivation

Visual analysis of similarities and contrasts in a data set can help analysts to monitor, explore and make sense of information more easily, particularly when little priori knowledge about the data is available except some initial hypotheses. Many real-world infographics and visualizations are created based on charts, such as line charts, area charts, bar charts and pie charts, that are familiar to most people. When multiple instances of charts are displayed together, however, it can be very difficult to identify the trend or compare the data unless they are well organized.

Small multiples, described by Jacques Bertin and popularized by Edward Tufte, allow one to examine multiple facets of a complex data set, and support visual comparisons and tracking of dynamic objects [12, 143]. They have been applied to monitoring and analyzing data-intensive processes such as system management, quality control, medical record analysis, and large-scale industrial and engineering operations. Small multiples are particularly useful for side-by-side visual comparison of multiple items without overplotting or occlusion that may occur when too many items are plotted together. However, as the number and complexity of the items increases, the effectiveness of small multiples quickly diminishes. This is because the amount of work required to examine the charts one by one

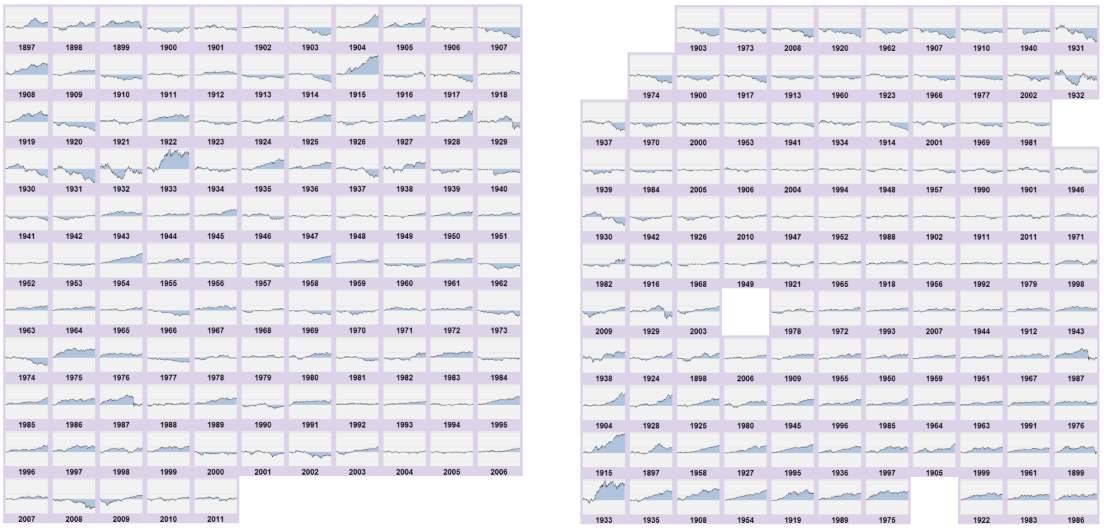


Figure 3.1: The Dow Jones Industrial Average (DJIA) from 1897 to 2011. Left: rendered as sequential small multiples, ordered by year. Right: CorrelatedMultiples, in a spatially coherent layout based on similarity. Charts of the years 2008 and 1920 are similar and are placed close to each other at the top of CorrelatedMultiples (right), but are far apart in the sequential small multiples (left).

in order to obtain an understanding of the whole data set is very high, making the visualization not much better than reading a table [51]. Figure 3.1 (left) shows an example of small multiples of the Dow Jones stock market index from 1897 to 2011. From the visualization, it is hard to identify the most common trends and similar years from over a hundred years of history. To amend this, it is necessary to have a better way to display the small multiples and organize the individual plots based on correlations in the underlying data.

In this chapter[1], we present *CorrelatedMultiples*, a spatially coherent visualization based on small multiples. The goal of CorrelatedMultiples is to exploit similarity among a set of items to determine their layout. By incorporating similarity into the spatial layout,

[1]Major portions of this chapter were previously published in Liu et al. [90].

viewers can better judge the overall qualitative characteristics of the data set [51]. Following the idea of standard multidimensional scaling (MDS) [84], we model the relationships among data items as a similarity graph, and embed the items in the layout so that proximity reflects similarity. Similarity between items is measured by domain-specific metrics. To lay out the correlated items so that they fit within a fixed display space, we propose a new optimization algorithm, based on Constrained Multi-Dimensional Scaling (CMDS). We show that this optimization approach is **computationally efficient**, **visually stable** and **space-efficient** through numerical and visual comparisons with other state-of-the-art methods. In addition, it is also **easy to implement** the algorithm with an iterative solver. We conducted a controlled user study to validate the effectiveness of CorrelatedMultiples in searching similar items. We also demonstrate its usefulness in three domains: stock market trends, census demographics, and climate modeling.

CorrelatedMultiples allows users to identify similarities and differences in the data more effectively, because related items are placed nearby, and unrelated items are pushed farther away from each another. Aggregates and anomalies can be discovered and examined based on the items' spatial locations, since adjacent items are likely to be similar, making abnormal events easier to identify. The main contributions of this work are:

- We introduce CorrelatedMultiples, which encode data similarity using spatial proximity among items in small multiples.
- We propose a new optimization algorithm, based on constrained multidimensional scaling, to create CorrelatedMultiples with both computational efficiency and visual stability.

- We evaluate the effectiveness of CorrelatedMultiples through a controlled user study, and three real-world case studies.

3.2 From Small Multiples to CorrelatedMultiples

Small multiples have been used widely for visual comparison of multiple items and alleviating overplotting or occlusion that may occur if multiple data sets are plotted in a single visualization. A common visual task is to find items that are very different from most others (anomalies), while another type of task is to find similar items that suggest interesting patterns for further analysis. Visual search is a perceptual task requiring a user's attention and typically involves an active scan for a particular object (the target) among other objects (distractors). Compared with the search features provided by many text-oriented information retrieval systems, visual search is particularly useful when the user has little knowledge about the structure of the data. Practical examples can be seen in our everyday life, such as looking for fashionable clothes while shopping, or detecting an oddly-behaved person during security monitoring. For small multiples, as the number and complexity of items increase, the resulting large visual search space reduces the efficiency of visual comparison, because it requires users to scan through many items with few visual cues.

The idea behind CorrelatedMultiples is to retain the key advantages of small multiples: (1) displaying multiple instances of data with a consistent representation, allowing side-by-side item comparison; (2) avoiding occlusion and visual clutter; (3) making effective use of display space compared with non-space-filling visualizations such as graphs and maps. In addition, CorrelatedMultiples enhance small multiples by (4) reflecting item-item similarity in spatial proximity. The goal of CorrelatedMultiples is to reduce users' visual

search space when searching for similar items near the target, dissimilar items far from it, or anomalies at the periphery of the layout. Our hypothesis is that such a spatial arrangement improves the performance of visual comparison tasks. This was tested in a controlled user study presented in Section 3.4.1.

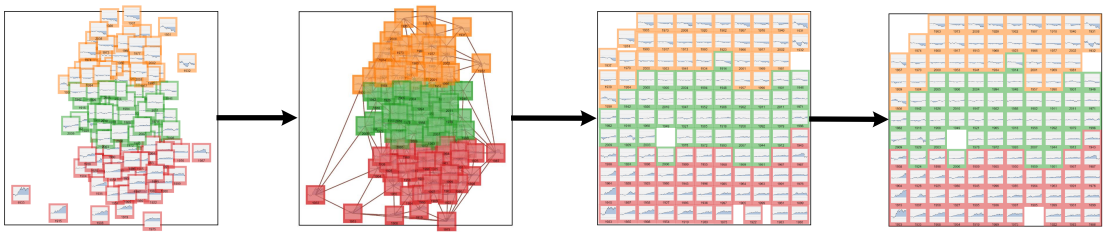
The main challenge addressed in this dissertation is to allow plotting spatially coherent small multiples in a constrained space with computational efficiency, visual stability and space-efficiency. To tackle this challenge, we first model the relationships among small multiples as a similarity graph. The similarity between two items is measured by some appropriate domain-specific distance function for the underlying data. To generate CorrelatedMultiples, we first assign initial coordinates of items in the display area using multidimensional scaling so that spatial proximity reflects similarity (Figure 3.2(a)). Next, a proximity graph is derived from a Delaunay triangulation of the items (Figure 3.2(b)). To lay out the correlated items within the available display space while maintaining spatial proximity (Figure 3.2(c)), we propose a Constrained Multi-Dimensional Scaling (CMDS) model described in Section 3.3. Also, as a post-processing step, items are aligned horizontally and vertically to maintain the clarity of the overall appearance (Figure 3.2(d)).

3.3 Constrained Multidimensional Scaling Algorithm

In this section, we describe the Constrained Multi-Dimensional Scaling (CMDS) algorithm for CorrelatedMultiples. The proposed algorithm is also applicable to general graph visualization.

3.3.1 Model Formulation

In graph drawing, stress based multidimensional scaling has been applied to generate high quality layouts to approximate predefined target distance between objects. The full



(a) Initial Placement (b) Proximity Graph (c) Layout using CMDS (d) Layout in Alignment

Figure 3.2: Overview of the CorrelatedMultiples pipeline: (a) placing data items based on similarity; (b) computing a proximity graph by Delaunay triangulation to constrain relative positions of items; (c) adjusting the layout by the proposed CMDS algorithm; (d) aligning the final layout horizontally and vertically.

stress model assumes there are springs between all node pairs. Given a 2-D layout where node i is placed at point p_i, the energy of the spring system is

$$\sum_{\{i,j\}\in V} w_{ij}(\|p_i - p_j\| - d_{ij})^2, \tag{3.1}$$

where d_{ij} is the ideal distance to be achieved between nodes i and j, and w_{ij} is a weighting factor, typically $1/d_{ij}^2$. A layout that minimizes the total stress is considered optimal. We formulate CMDS from this basic stress model.

Given an initial layout of the items, we first generate a proximity graph $G = (V, E)$, with V the set of nodes (items) and E the set of edges. The proximity graph links items that are close by in the initial layout with edges. Maintaining the length of these edges as much as possible helps to preserve the proximity relation between items. With that in mind, let the display region be Γ, then our goal is to find new coordinates p_i for each node $i \in V$, such that: (1) there is no overlap between any nodes $\{i, j\} \in V$; (2) each edge $(i, j) \in E$ is close to its ideal length; (3) each p_i is inside Γ. The first condition, no-overlap, is needed for readability and aesthetics; the second preserves spatial proximity of nodes in the graph. At the same time, the graph should fit in Γ to utilize the available space without being scaled

down so much that it reduces readability. In small multiples, nodes are not just points but have the same finite area, so the Γ must be shrunk by the margin of half the node's width and height, so that every node can be fully displayed as long as its center lies within Γ. With these conditions in mind, we propose the CMDS model

$$\min \sum_{(i,j)\in E} w_{ij}(\|p_i - p_j\| - d_{ij})^2 + \alpha \sum_{p_i \notin \Gamma} (\|p_i - \Gamma(p_i)\|)^2, \tag{3.2}$$

where $\alpha \geq 0$ is a parameter to balance the two terms, and $\Gamma(p_i)$ denotes the projection of p_i to the region Γ — if p_i is outside of Γ, the projection $\Gamma(p_i)$ gives the point on the boundary of Γ that is closest to p_i; otherwise $\Gamma(p_i) = p_i$.

The first term in (3.2) encodes the stress energy between nodes sharing an edge. The initial value of p_i is determined by similarity-based embedding (such as standard MDS), or may be supplied by the user. The initial distance between node i and j is denoted as l_{ij}. Subsequently, to maintain the proximity relation, we follow the PRISM approach [38]: we take a Delaunay triangulation of the layout, set this to be the proximity graph G (thus, E is the set of triangulation edges), and solve the optimization problem to preserve node distances along triangulation edges. The rigidity of the triangulation provides sufficient scaffolding to constrain the relative positions of the components, and helps preserve the global structure of the original embedding. Fortunately, we only need to consider pairs of nodes that share an edge, which yields a sparse model that can be solved much faster than a full stress model in (3.1). For each edge (i, j), the amount of overlap on line $p_i \rightarrow p_j$ is denoted by δ_{ij} (δ_{ij} is set to 0 if no collision is detected). Since our goal is to remove overlap δ_{ij} while preserving the initial distance l_{ij} as much as possible, the ideal distance is set to $d_{ij} = l_{ij} + \delta_{ij}$. The second term of (3.2) is the stress energy between node i and Γ if node i is outside of Γ. The ideal distance between node i and its projected point $\Gamma(p_i)$ is 0.

Taking the gradient of (3.2) with respect to p_i, assuming that $\Gamma(p_i)$ is constant, and setting the gradient to zero gives

$$\sum_{(i,j)\in E} w_{ij}(\|p_i - p_j\| - d_{ij})\frac{p_i - p_j}{\|p_i - p_j\|} + \alpha(p_i - \Gamma(p_i)) = 0, \tag{3.3}$$

By algebraic manipulation of (3.3), keeping linear terms involving p_i to the left and moving the rest to the right of the equation, we obtain the following iterative scheme

$$p_i(t+1) \leftarrow \frac{\sum_{(i,j)\in E} w_{ij}\left(p_j(t) + d_{ij}\frac{p_i(t)-p_j(t)}{\|p_i(t)-p_j(t)\|}\right) + \alpha\Gamma(p_i(t))}{\sum_{(i,j)\in E} w_{ij} \; + \; \alpha}. \tag{3.4}$$

This iterative process does not require the solution of a linear system. This makes it easier to implement in browser supported languages such as JavaScript that do not have sophisticated numerical libraries. More importantly, by rendering the objects during the iterations, we can visualize the change of the node positions, from an initial unconstrained configuration, to the final constrained layout. This is shown in the video that accompanies this dissertation.

After N iterations of (3.4), the layout may still have overlaps and nodes outside display space. If so, we regenerate the proximity graph using a Delaunay triangulation augmented with additional edges for overlapping nodes, calculate ideal edge lengths, and then return to the energy minimization step. We refer to this algorithm as the Constrained Multi-Dimensional Scaling (CMDS) algorithm, shown in Algorithm 1.

Time Complexity. If the initial embedding is done through approximate MDS, it can take $O(|V|\log|V|)$ time [69], or $O(1)$ if it is given by the user. A Delaunay triangulation can be found in $O(|V|log|V|)$ time. A scan-line algorithm to find all the overlaps can be implemented in $O(l|V|(\log|V|+l))$ time [29], where l is the number of overlaps. A sweep-line algorithm has a similar time complexity ($O((l+|V|)\log|V|)$) [26]. Because we only apply the scan-line algorithm after all node overlaps are removed along proximity graph edges, l is usually a very small number, hence this step can be considered as having complexity

Algorithm 1 Constrained Multi-Dimensional Scaling

Input: the coordinates p_i of data items; the region Γ.

```
Construct a proximity graph G by Delaunay triangulation.
repeat
    Calculate the ideal distance for all edges.
    while (iteration < N) do
        Update each p_i according to (3.4).
    end while
    Construct a proximity graph G by Delaunay triangulation.
    Augment G with edges from pairs of nodes that overlap.
until (All data items are inside Γ with no overlap)
```

$O(|V|\log|V|)$. Calculating ideal distances takes $O(|E|)$ time, and iteratively solving the proposed model takes $O(N|E|)$ time when edges for each node are pre-stored. Therefore, overall, Algorithm 1 takes $O(c(|V|log|V|+|E|+N|E|))$ time, where c is the total number of iterations in the outer loop of Algorithm 1. In our implementation (described in the next sub section), $N = 20$ is generally found to achieve good layouts.

Convergence. Given sufficient display space, as Algorithm 1 runs, eventually $\Gamma(p_i) = p_i$ and the second term of (3.2) disappears. The convergence of the first term (denoted as $S(p)$) follows from the convergence of stress majorization [39], and thus the CMDS model converges. To achieve this, the input nodes are uniformly scaled so that the total sum of their area is sufficiently less than Γ. We now theoretically prove that the stress of the CMDS model converges if the display boundary is a circle. The case where the display boundary is not a circle is more complex due to the potential non-smoothness of the boundary. However empirically, as we shall see in the next subsection, it also converges for rectangular display space.

For a circular display space Γ with radius r, $\Gamma(p_i) = rp_i/\|p_i\|$ assuming the circle center is the origin, the total stress in (3.2) can be rewritten as

$$\begin{aligned} Stress(p) &= S(p) + \alpha \sum_{p_i \notin \Gamma} (\|p_i - \frac{p_i}{\|p_i\|} r\|)^2 \\ &= S(p) + \alpha \sum_{p_i \notin \Gamma} (\|p_i\| - r)^2 \\ &= S(p) + \alpha \sum_{p_i \notin \Gamma} (\|p_i\|^2 - 2\|p_i\| r + r^2) \\ &\leq S(p) + \alpha \sum_{p_i \notin \Gamma} (\|p_i\|^2 - 2r p_i^T q_i/\|q_i\| + r^2) = F^q(p), \end{aligned}$$

with equality when $q = p$ (following the Cauchy-Schwartz inequality). This way we have bounded the stress with a quadratic form $F^q(p)$. Given some layout $p(t)$, if we set $q = p(t)$ and minimize the quadratic function $F^q(p)$ to get $p(t+1)$, then we know that

$$Stress(p(t+1)) \leq F^q(p(t+1)) \leq F^q(p(t)) = Stress(p(t)).$$

The first inequality is due to the bound above. Thus the stress of the sequence $p(t)$ decreases monotonically with regard to t. Therefore, the stress will eventually converge to a local minimum. We differentiate $F^q(p)$ by p and the minima are given by solving

$$(L^w + \alpha I)p = L^q q + R^q, \tag{3.5}$$

where L^w (weighted Laplacian) and L^q are defined as

$$L^w_{i,j} = \begin{cases} -w_{ij} & i \neq j \\ \sum_{k \neq i} w_{ik} & i = j \end{cases}, L^q_{i,j} = \begin{cases} -w_{ij} d_{ij}/\|q_i - q_j\| & i \neq j \\ -\sum_{k \neq i} L^q_{i,k} & i = j \end{cases},$$

and $R^q = \alpha \sum_{p_i \notin \Gamma} r q_i/\|q_i\|$, and I is the identity matrix. Therefore, the above majorization process requires solving $(L^w + \alpha I)p(t+1) = L^{p(t)} p(t) + R^{p(t)}$ until the stress converges, which can be expressed using the iterative scheme in (3.4).

Grid Alignment. To arrange the multiples in a grid layout like conventional small multiples, the final CMDS layout is aligned horizontally and vertically using a *round* function:

$$\begin{cases} x_a = round((x - w/2)/w) \times w + w/2 \\ \\ y_a = round((y - h/2)/h) \times h + h/2 \end{cases}, \tag{3.6}$$

where x and y are the coordinates of the center of a node, w and h are the width and height of the node, x_a and y_a are the aligned center of the node. Essentially, the round function aligns a node to the grid cell that covers the center of the node. Given that the area of a node is equal to that of a grid cell, it is clear that if nodes do not overlap, one grid cell can cover at most one node, and thus this simple round function is able to produce a grid-like layout without overlap after alignment (as shown from Figure 3.2(c) to (d)).

3.3.2 Parameter Study

Given a display space $\Gamma = W \times H$, an $m \times n$ grid that fits Γ is determined such that it can sufficiently cover the N nodes ($N \leq m \times n, m \times (n-1) \leq N, (m-1) \times n \leq N$). The size of a grid cell (also the size of a node) is then set to $\frac{W}{m} \times \frac{H}{n}$. If there is still overlap after a number of iterations using CMDS, either m or n will be increased by 1 and the nodes will be rescaled. We found that this heuristic approach can ensure that the total sum of area of nodes is sufficiently less than Γ for convergence. In case that the visualization in each node (e.g., an area chart or a bar chart) has special aspect ratio constraint, each node can be adjusted to fit its own aspect ratio within the bound of a grid cell as a post-processing step.

Several additional parameters need to be defined in the implementation, such as the number of iterations N in Algorithm 1, and the parameter α in (3.4). Our approach is to study the trade-off (Pareto curve) between stress and display-space utilization, and choose

parameters that yield the best results. We first define the area loss measure (AL)

$$AL = 1 - \frac{\sum_{i \in V} A_i}{B_V + t * \sum_{\{i,j\} \in V} A_i \cap A_j}, \tag{3.7}$$

where A_i is the area that node i covers, B_V is the bounding box of all nodes, and t is a penalty factor for node overlaps. This penalty must be accounted for, because otherwise a layout where all nodes placed on the same point would waste the least area. Since overlap removal is essential for small multiples, we set $t = 10$ after testing different values experimentally. The stress with reference to the initial structure is denoted as

$$Stress = \sum_{(i,j) \in E} w_{ij}(s\|p_i - p_j\| - l_{ij})^2, \tag{3.8}$$

where w_{ij} is a weighting factor, l_{ij} is the initial length of edge (i, j), and s is a scaling factor that minimizes (3.8)

$$s = \frac{\sum_{(i,j) \in E} w_{ij} l_{ij} \|p_i - p_j\|}{\sum_{(i,j) \in E} w_{ij} \|p_i - p_j\|^2}. \tag{3.9}$$

For each parameter, our goal is to find a value that achieves good space usage while keeping the stress low. Given an initial placement, we run CMDS and calculate area loss (3.7) and stress (3.8) for different settings of the number of iterations N in Algorithm 1 and α in (3.4). It was observed that CMDS converges with varying stress and area loss for different parameters (the ending location in Figure 3.3). As shown in Figure 3.3 (top), when the number of inner iterations N is too small ($N = 1$), space utilization is poor. This is because, without enough iterations of (3.4), the ideal distance is updated based on a configuration that has not yet been reached. Too many inner iterations ($N \geq 40$) might result in higher stress because of doing too much work to satisfy ideal edge length that were determined before the inner iteration starts, and most of the overlaps are usually removed after a few iterations. Excessive iterations without rechecking overlap generally makes layouts that deviate too much from the initial layout, increasing the stress. Therefore, based

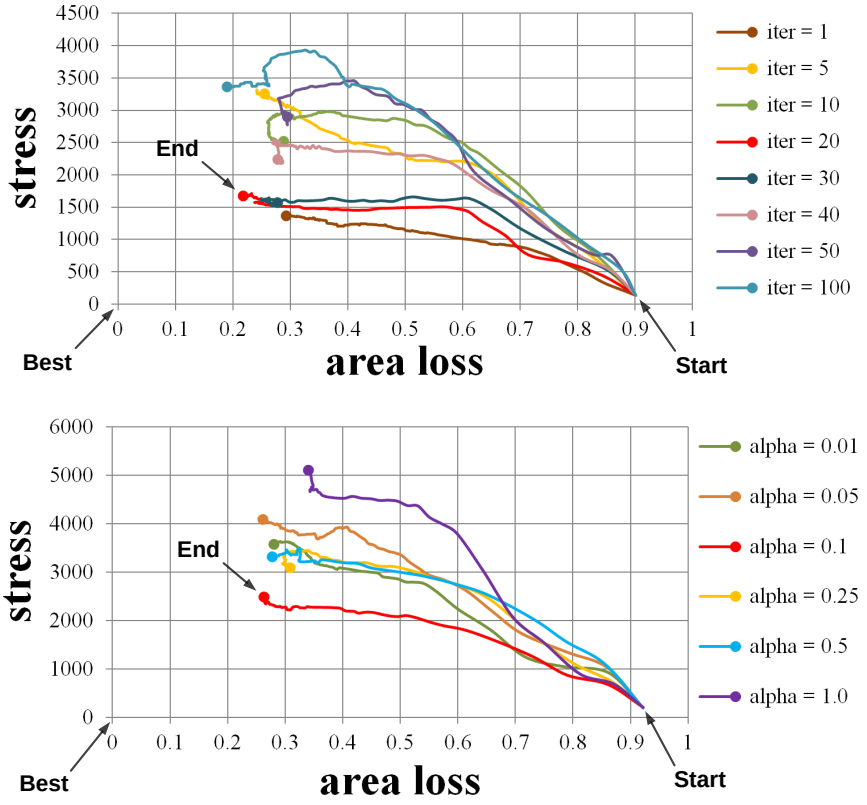


Figure 3.3: Stress-AL Pareto curves showing the progression of iterations applied to 100 nodes in a 1100×1100 pixel screen space: (top) for the number of inner iterations N in Algorithm 1, too large or too small values will lead to high stress or poor space utilization; (bottom) for α in (3.4), large or small values will result in high stress or poor space utilization.

on Figure 3.3 (top), we set $N = 20$ to balance stress reduction with area utilization. Figure 3.3 (bottom) demonstrates the effect of α in (3.4) with the Pareto curve of stress and area loss. If α is too small ($\alpha = 0.01$), we are essentially solving a conventional sparse stress model without the boundary constraint; on the other hand, if α is large ($\alpha \geq 0.25$), flipping may occur because of abrupt changes of nodes close to the boundary. Experimentally, we found that $\alpha = 0.1$ yields good results.

3.4 Evaluation

This section presents several studies to evaluate the effectiveness of CorrelatedMultiples. The user study (Section 3.4.1) is focused on tasks related to visual search for similar items, where we compared CorrelatedMultiples with sequential small multiples and an existing grid layout method. The quantitative study (Section 3.4.2) shows the performance of the CMDS algorithm in comparison with several competing methods. We also describe three case studies (Section 3.4.3) using different real-world datasets to demonstrate the usefulness of CorrelatedMultiples.

3.4.1 User Study

Recently, Gomez-Nieto et al. [46] have found that 2D similarity-based visualization allows users to identify related items faster than using 1D ranked list with comparable accuracy. In the context of image browsing, Rodden et al. [119] have shown that users were faster to find a photograph when similar images were arranged close to each other than a random layout. However, user performance on comparing multiple abstract graphical depictions such as line charts or area charts has not been studied in the literature. To evaluate how much a spatially coherent layout affects task performance in visual search for similar items that contain abstract graphical visualization, in this study we compare sequential small multiples (denoted as **SM**), the spiral-search-based grids by Rodden et al. [119] (denoted as **SG**), and CorrelatedMultiples (denoted as **CM**).

Experiment Setup

We collected CPU usage from 5144 network devices, sampled at 5-minute intervals over one day. We plotted the time series of 288 data points for each device using an area

chart, and randomly selected a subset of the charts to make three types of visualizations under study. The charts in **SM** were arranged by device ID in ascending order, while those in **SG** and **CM** are based on data similarity, which is measured by *Dynamic Time Warping (DTW)*, a technique that finds a warping path which minimizes the total distance between two series. Among the three heuristics proposed by Rodden et al. [119], we took the bumping strategy in generating **SG** as it is overall the best.

We recruited 21 subjects (14 males, 7 females) having various backgrounds in computer science, software engineering, electrical engineering, chemical engineering, geographic information science, physics, chemistry, accounting, economics, and finance. The subjects ranged in age from 23 to 32 years old, with a mean age of 26. All of the subjects said they were unfamiliar with the notion of small multiples.

Subjects were asked to perform a visual search task — identify one or multiple items that are the most similar to a target given in the visualization, which involves visual comparison [46].

Procedure

The study was conducted as a within-subjects experiment with 3 experimental conditions (**SM**, **SG**, and **CM**) and 9 repetitions (visualization image with varying number of charts (from 50 to 150)) for each condition. For each repetition, the subject was presented with only one condition. We counter-balanced the selection of conditions in the 9 repetitions so that each subject performed one repetition for all three conditions with the same number of charts.

The study was performed on an i7-2600, 3.4 GHz CPU desktop computer with a standard 24-inch screen of 1920 $\times$ 1080 pixels. Prior to the experiment, subjects viewed a tutorial that has a basic explanation of small multiples, and then performed some training

tasks to get familiar with the user interface of the experimental system. For each task, the subject was given a randomly chosen **SM/SG/CM**, and prompted to answer the question: *Find "i" most similar area chart(s) to the highlighted one* ($i = 1$ or 2 randomly). They click on the "next" button to load the next task. After the subjects finished all tasks, they were asked to rate their satisfaction with **SM/SG/CM**, on a questionnaire containing 3 questions, and finally, to participate in a semi-structured interview.

Results and Discussion

We measured the task completion time that the subjects needed to find items similar to the given target, the accuracy of their selections, and the subjective assessment from questionnaires. Task completion time and accuracy measures were evaluated using single factor Analysis of Variance (ANOVA) for the dependent variables. A logarithmic transformation was applied to the task completion time as a standard approach to correct the non-normal distribution of such time data. A significant effect of visualization type was found on task completion time ($F(2,60) = 4.8776, p < 0.05$). The average time spent on a task was 23.15 seconds for **SM**, 17.46 seconds for **SG**, and 13.81 seconds for **CM**, as illustrated in Table 3.1. Task accuracy was not found to have a significant difference among the three visualization types ($F(2,60) = 2.4053, p = 0.0989$), but only between **SM** and **CM** ($F(1,40) = 4.5924, p < 0.05$). On average, the accuracy of a task was 63.5% for **SM**, 73.0% for **SG** and 84.1% for **CM**.

The questionnaire asked subjects to assess their satisfaction with **CM/SG/SM** on multiple criteria: spatial proximity (*C1: SM/SG/CM is the best to place similar charts nearby and dissimilar ones far apart*), efficiency (*C2: I could find similar items more quickly with SM/SG/CM*), and confidence (*C3: I was more confident in my answer with SM/SG/CM*). It was found that subjects were more efficient and confident in the visual search tasks with

Table 3.1: Results of the user study of CorrelatedMultiples.

	Time	Accuracy	C1	C2	C3
SM	23.15s	63.5%	0%	4.8%	9.6%
SG	17.46s	73.0%	33.3%	19.0%	19.0%
CM	13.81s	84.1%	66.7%	76.2%	71.4%

CM than with **SG**, and least with **SM**. Two thirds of them viewed **SM** as spatially more coherent than **SG**. In the interviews, three subjects mentioned that it was a ”cool” or ”great” idea to place similar items together rather than randomly spread them out in the display. One subject said searching a potential candidate in the visualization with 150 small charts was frustrating if the items were not arranged in a spatially coherent layout. The overall feedback on **CM** obtained in the interviews was positive. Ten subjects mentioned that **CM** was ”helpful”, ”beneficial” or ”useful”.

In summary, encoding spatial coherency within small multiples was valued by the users. Spatially coherent layouts are useful in helping users visually identify similar items, and they can perform this more quickly and more accurately with CorrelatedMultiples in which the spatial proximity is well preserved. Most users preferred CorrelatedMultiples in visual search for similar graphical charts.

3.4.2 Quantitative Study

Space-Efficiency. The goal of the CMDS algorithm for generating CorrelatedMultiples is closely related to overlap removal in the field of data visualization. The main difference of CMDS from most existing methods is that it considers the boundary of display space as an important constraint to generate a space-efficient layout. To show this, we compared

the CMDS algorithm with the PRISM algorithm [38], which is the closest to our work. Figure 3.4 shows an initial layout (left) and the layout after overlap removal using PRISM (right). The layout after removing overlaps of Figure 3.4 (left) using CMDS is shown in Figure 3.1 (right). It can be observed that without the boundary constraint, many items went out of the display boundary after overlap removal using PRISM, which leads to loss of information. Although the items can be uniformly scaled down to fit the display space, it can lead to readability issue and inefficient space utilization. In contrast, our CMDS algorithm avoided such undesirable outcomes by constraining all items within the display.

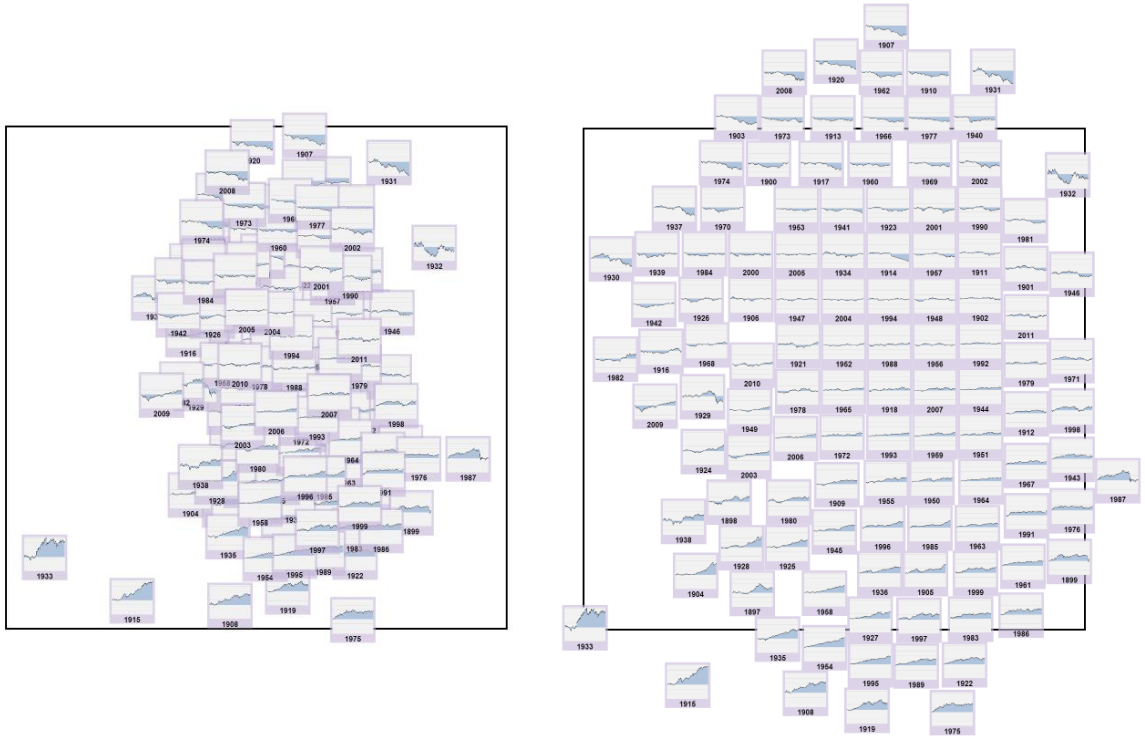


Figure 3.4: Overlap removal using PRISM [38]. Left: an initial layout of the stock market charts. Right: a layout after overlap removal using PRISM. The layout created by CMDS is shown in Figure 3.1.

Computational Efficiency and Visual Stability. The user study presented in Section 3.4.1 implies that a layout with good spatial proximity can improve user performance in searching similar items. To further investigate the performance of the CMDS algorithm, we conducted a quantitative experiment comparing CMDS with six methods for producing spatially coherent grid layouts. Besides the *BumpSpiralGrid* method that has been used in the user study (i.e., **SG** in Section 3.4.1), we further include *BasicSpiralGrid* and *SwapSpiralGrid* (also by Rodden et al. [119] but with two other greedy heuristics), the gridded spatially ordered Treemap by Wood and Jason [164] (denoted as *SpatialTreeMap*), the method by Vaidya et al. [151] that minimizes displacement without translation or scaling (denoted as *SimpleGridMap*), and the method detailed by Eppstein et al. [32] that minimizes the displacement with translations (denoted as *TransGridMap*) [1]. The first four are basically greedy approaches, while the last two as well as CMDS are optimization methods.

To evaluate the visual stability, we employed a set of metrics derived from [164] and [32] — the ratio of displacement (*disp.*), the percentage of recalled adjacency (*recall*) and the percentage of the preserved directional relation (*presv.*). We also measured the computational time performance (CPU) of each method. All methods were implemented in Javascript and HTML5. The optimal point set matching for SimpleGridMap and TransGridMap was achieved using integer linear programming (ILP) based on a GNU Linear Programming Kit for Javascript (GLPKJS) [44]. We ran each method on a geographical map of United States (from Eppstein *et al* [32]) in a Google Chrome browser on an i7-2600, 3.4 GHz CPU computer. To be consistent with the experiment by Eppstein et al. [32], we

[1] SimpleGridMap and TransGridMap are known as the I and L_1 methods in [32]. Two other methods, W and L_2^2, are discussed in [32]. W is a 4-approximation of the problem of minimizing directional reversal, by solving an L_1 optimization with weights defined by rank differences. L_2^2 is a method proposed by Cohen and Guibas [22]. Both have comparable quality to L_1, but W has a complexity $O(n^2 log^3 n)$ as opposed to $O(n^6 log^3 n)$ for L_1. We choose to compare with I and L_1 because I is among the fastest method of the 4 discussed in [32], and L_1 is representative in quality to the best methods known.

Table 3.2: Numerical measures of seven grid layout methods, measuring displacement (disp.), recalled adjacency (recall), preserved directional relation (presv.) and CPU time.

Method	disp.	recall	presv.	time
TransGridMap	0.1767	74.29%	90.69%	6.67h
CMDS	0.1811	75.24%	89.80%	0.162s
SimpleGridMap	0.2508	75.24%	89.72%	0.521s
SpatialTreeMap	0.2697	73.33%	89.01%	0.010s
BumpSpiralGrid	0.3372	29.81%	72.70%	0.022s
SwapSpiralGrid	0.3632	38.46%	69.15%	0.020s
BasicSpiralGrid	0.3754	42.78%	72.34%	0.018s

considered only the 48 contiguous states as the initial layout, and the target layout as an 8×6 grid.

Table 3.2 shows the results of this experiment, with the methods sorted by displacement in ascending order. Comparing the displacement and time performance, it can be seen that the top three optimization methods resulted in less displacement at varying costs in terms of the computation time: TransGridMap performed best in minimizing displacement (which was not a surprise since the method is the best known optimization algorithm in minimizing this quantity), but took an overwhelming amount of time (since it requires the computation of $48 \times 48 \times 48$ distance matchings); CMDS yielded displacement close to TransGridMap (only 2.4% difference), but about 150,000 times faster than TransGridMap; in addition, CMDS outperformed the other methods in minimizing displacement — SimpleGridMap (27.8% improvement and also 3 times faster), SpatialTreeMap (32.9% improvement), and the three SpiralGrid methods (over 40% improvement for each); on the other hand, the bottom four greedy methods were more computationally efficient (about 10 times faster than

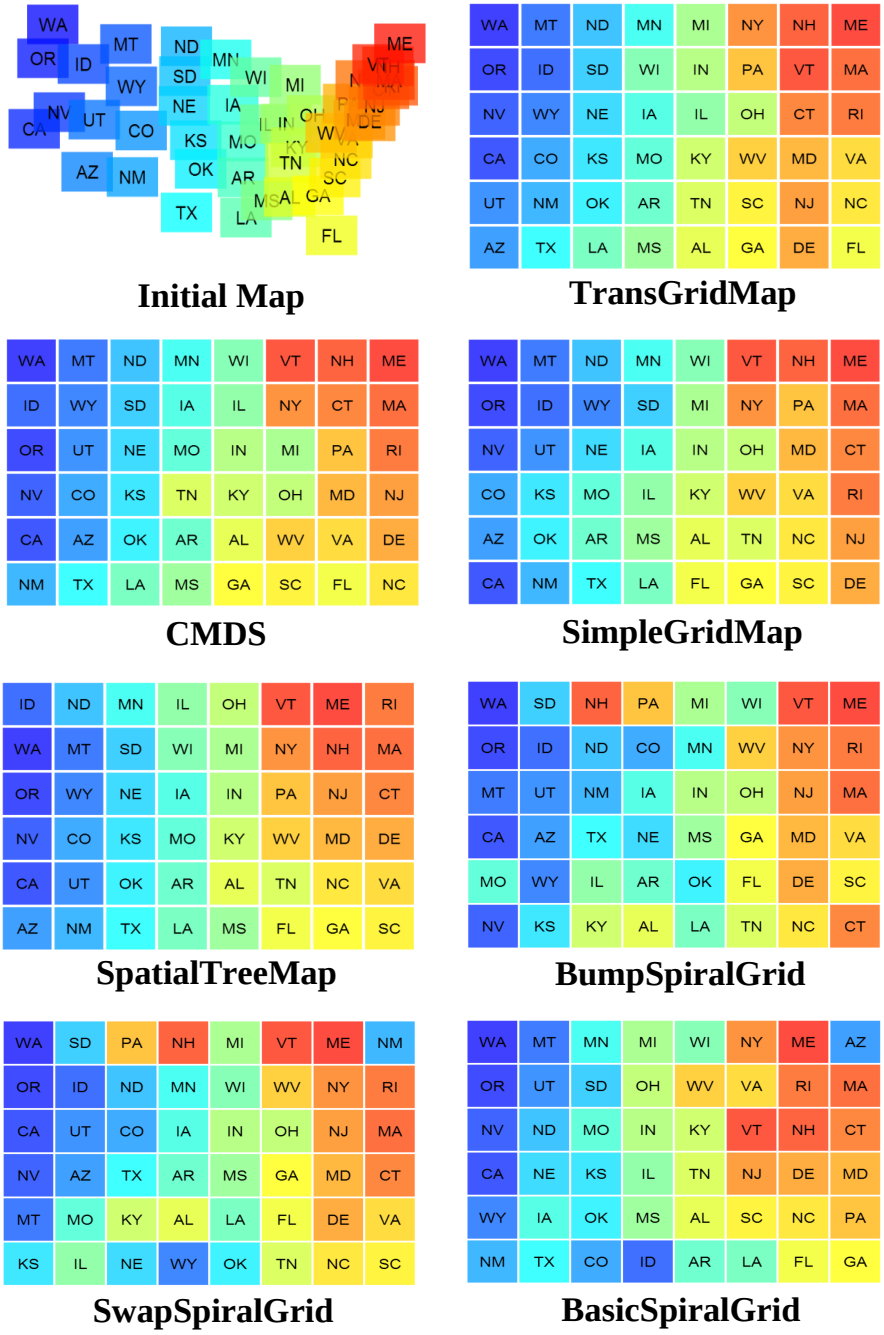


Figure 3.5: Various grid layouts of a USA map that consists of the 48 contiguous states.

CMDS and 25 times faster than SimpleGridMap). In terms of the recalled adjacency and preserved directional relation, the top four methods were almost equally good and better than the three SpiralGrid methods. This is partly because the spiral-search-based methods produced a lot more errors in dense grids than in sparse grids (shown in the previous study by Rodden et al. [119]).

The corresponding grid layouts in Figure 3.5 further explained these quantitative results. It is obvious that quite a few cells in the three SpiralGrid layouts were far away from their original positions, such as NH, PA, MO in BumpSpiralGrid, PA, NH, NM in SwapSpiralGrid, and AZ, OH, ID in BasicSpiralGrid. Some cells in SpatialTreeMap were placed relatively far from their original positions. For instance, the group of cells IL, OH, WI, MI, which were located in the middle right of the initial map, were placed in the middle top of SpatialTreeMap. As for SimpleTransGrid, cells such as FL and CO were placed relatively far from their original positions. Cells in TransGridMap and the CMDS layout are relatively stable compared with the other methods.

Overall, from both the numerical and visual comparisons, CMDS achieved a good balance between computational efficiency and visual stability in generating a spatially coherent layout.

3.4.3 Case Studies

We studied the effectiveness of CorrelatedMultiples in three real-world datasets of different fields: stock market data, census demographics data and climate simulation data.

Stock Market Data

Firstly, we studied the Dow Jones Industrial Average (DJIA) trends by year, from 1897 to 2011. Each year contains about 250 time steps (all weekdays). Since we are interested

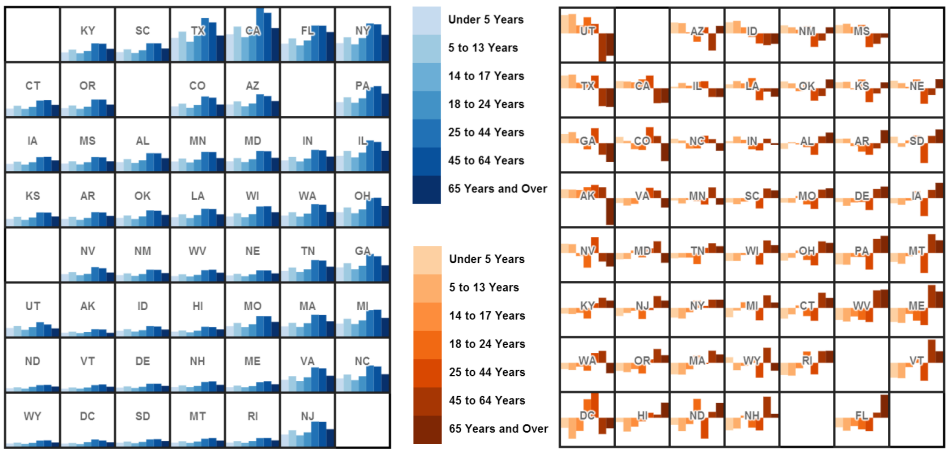


Figure 3.6: CorrelatedMultiples of the population charts (left) and variation charts (right) for the 2008 U.S. census data.

in relative fluctuations in the DJIA, for each year we placed the beginning of the year at the origin, and plot the percentage of change at each time step.

To measure similarity in the DJIA time series between two years, we use the same DTW-based similarity measure as in Section 3.4.1. Figure 3.1 shows two different visualizations of the DJIA over 115 years. In the sequential small multiples (left), charts were arranged in the ascending order of year, while the charts in CorrelatedMultiples (right) were placed in a spatially coherent manner.

In the CorrelatedMultiples of Figure 3.1 (right), we can see three main groups, which are falling, stable, and rising prices (from top to bottom). In general, the closer to the top in the figure, the more the DJIA fell over time in the corresponding year. Near the bottom of the figure, we see increases in the index over a year. As expected, stable trends are placed in the middle of the figure. Because of the spatial proximity that CorrelatedMultiples preserve, trends of similar fluctuations are placed nearby, which helps us identify

similar patterns. For example, most of the trends at the top of the figure in the CorrelatedMultiples reflect severe economic crises, such as in 2008, 1907 and 1931 (from left to right in the top row), all of which saw sharp declines (around 50% during the year). These correspond to *the Global Financial Crisis*, *the 1907 Bankers' Panic* and *the Great Depression*, respectively. Interestingly, a dramatic rising trend of the year 1933 in the bottom-left of the CorrelatedMultiples (which is the farthest from the year 1931 in the visualization) represents a dramatic rebound from the Great Depression — more than 75% of the nation's banks had been reopened and several acts were passed by the U.S. Congress to revive the economy in that year.

Census Demographics Data

Secondly, we experimented with CorrelatedMultiples to encode demographic similarities and contrasts across U.S. states up to July 1, 2008. For each state, we partitioned the data by age into 7 groups. Hereafter, we denote *population charts* as histograms that visualize the population values in various age groups, and *variation charts* as histograms that visualize the deviation of the population percentages of one specific age group from that in the overall census data. We measure the distance between two population charts by an average of Euclidean distance and Jensen-Shannon distance, and use Euclidean distance between variation charts.

Figure 3.6 (left) shows CorrelatedMultiples of population charts for the U.S. states. We can see that the four most highly populated states (TX, CA, FL and NY) were placed together at the top; the states with the lowest population (such as DC, ND and VT) were placed at the bottom; the others in between appeared roughly around the center. Furthermore, we can easily find the states with similar population distributions simply by checking

the nearby charts, such as FL and NY (at the top-right), as well as NM, WV and NE (in the center)).

While population charts allow basic comparisons, variation charts, shown in Figure 3.6 (right), help us see which states have individual age segments that deviate significantly from the norm. For example, we can see that UT (at the top-left) is highly populated by people under age 44 while population above 45 is far below the overall average. In contrast, FL (at the bottom-right) has a relatively greater elderly population and a smaller young population. Quite distinct from the above two states, DC (at the bottom-left) mainly consists of young and mature adults.

Climate Simulation Data

In another case study, we studied the climate simulation of Madden-Julian Oscillation (MJO), a phenomenon manifested as intra-annual weather fluctuation in the tropics. We applied CorrelatedMultiples to a MJO simulation performed by the Pacific Northwest National Laboratory. The data set made by this simulation consists of 479 time steps, recorded at 6 hour intervals (from Oct. 1, 2007 to Jan. 29, 2008). We used the time-varying water vapor intensity collected in the region of $[60°E - 150°E]$ over time as the input data, and sampled it every 4 time steps (24 hours) to obtain 119 time steps (days). For each time step, we computed the distribution of water vapor by longitude in the given region, and rendered the distribution as an area chart (x-axis for longitudes from west to east, y-axis for water vapor values).

Since scientists are generally interested in studying time-varying patterns of MJO, such as shifts in the peaks of the values, we normalized the water vapor distribution at each time step, and computed *Earth Mover's Distance (EMD)* to measure the similarity of the

distributions between every two time steps. We generated CorrelatedMultiples using this dissimilarity measure, and cluster the items using K-Means based on the similarity.

Figure 3.7 shows CorrelatedMultiples of water vapor distributions from the MJO simulation. Time steps in the same cluster have similar distributions. Since rainfall oscillation is reflected in the shape of water vapor intensity, we can observe some interesting patterns from the top-right to the bottom-left of Figure 3.7: in cluster **A**, peaks generally showed up on the right, which means tropical rainfall reached its max in the east; while in cluster **B**, most of the peaks appeared in the middle; in cluster **C**, more peaks appeared in the west and east; and in cluster **D**, most charts have multiple peaks, evenly distributed, which suggests the MJO occurred more frequently in the corresponding time steps. Furthermore, it is apparent from the figure that some nearby charts with a large time interval have similar distributions, such as day 45 and 104, and day 10 and 114 (highlighted in the figure). This seems to imply a periodic pattern as MJO progresses.

3.5 Discussion

While we applied CorrelatedMultiples to line graphs and bar charts in our studies, our technique can also works well with many other types of visualizations, such as scatterplots, pie charts, Treemaps, as long as similarity between items can be computed. Since CorrelatedMultiples encode data similarity using spatial proximity, they can be augmented with additional visual cues such as highlighting different clusters in distinct colors (see Figure 3.2 and Figure 3.7). Since our user evaluation focuses on studying the effectiveness of layouts, this color encoding was not used to avoid influencing the user's decision.

Another lesson we learned is that CorrelatedMultiples and even conventional small multiples are only able to display tens or hundreds of items legibly. They do not scale up to

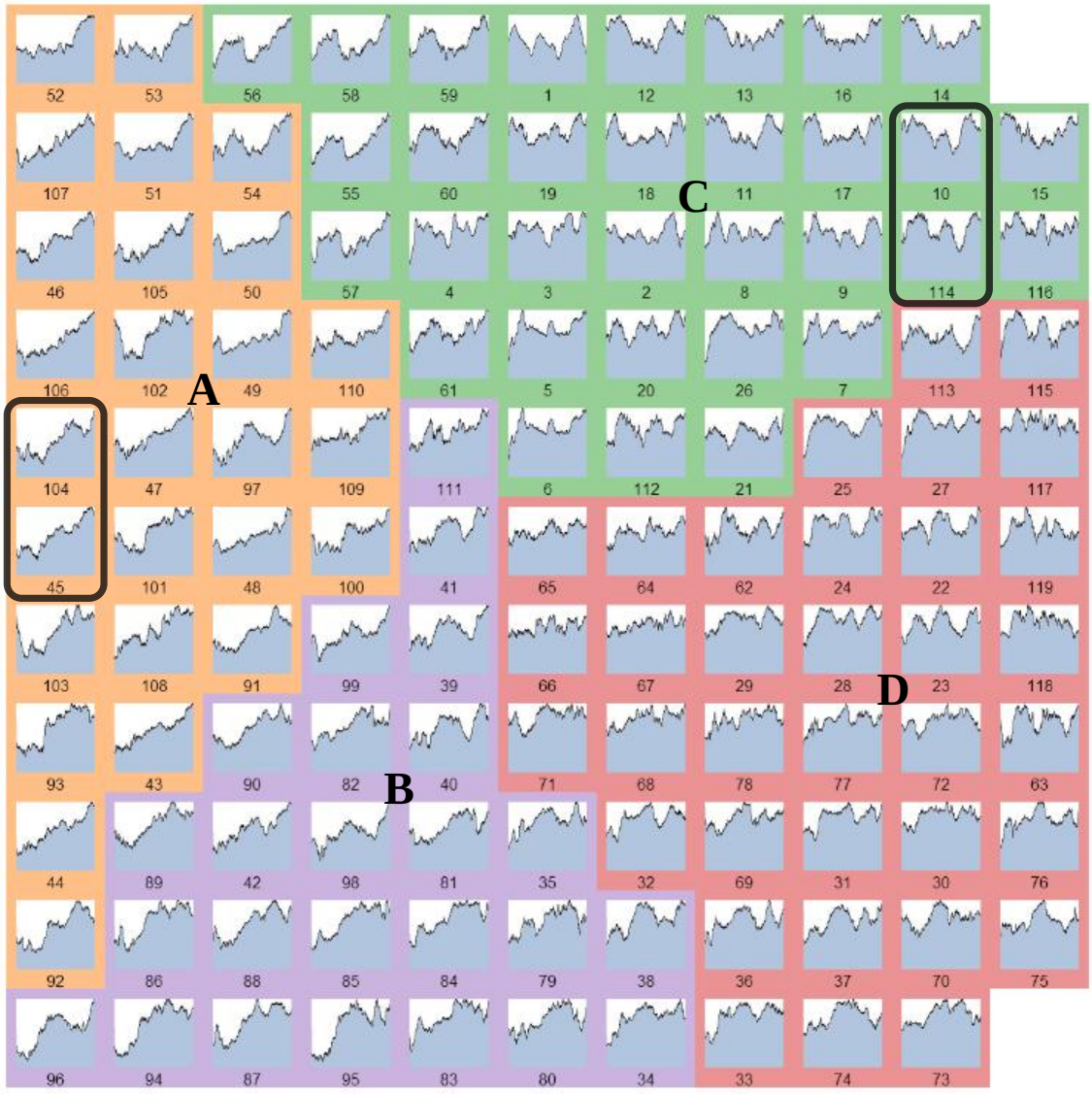


Figure 3.7: Correlated water vapor distributions in the Madden-Julian Oscillation (MJO) simulation visualized by CorrelatedMultiples.

thousands of items due to the space limit of conventional computer screens. This problem can be alleviated by enhancing CorrelatedMultiples with interactions such as zooming and panning or focus+context views to support details-on-demand visualization, or selecting a smaller but significant subset of multiples based on specific visualization tasks, which is

an active research problem in visual analytics of big data. Also, as the number of items increases, more area is wasted due to the trade-off of stress minimization and display-space utilization. The locations of the wasted empty cells can be random, but we believe such small wastage is offset by the benefits that CMDS brings, as we have demonstrated above.

Although in this study, CMDS was applied only to arranging small multiples that have the same finite areas on a grid, it has the potential for more general applications to produce overlap-free and space-efficient layouts in a non-grid layout. For instance, CMDS has been employed to create dynamic clustered visualization for simultaneously maintaining the user's mental map and maximizing screen space utilization [89]. In the future, we intend to adapt CMDS to applications where the display regions have a non-rectangular shape such as circular or user-specified form.

3.6 Summary

We proposed CorrelatedMultiples, a spatially coherent visualization for small multiples. CorrelatedMultiples retain the major advantages of small multiples, allowing side-by-side visual comparison, and enhance the visualization with spatial proximity based on similarity in the data. To generate CorrelatedMultiples, we presented an efficient and stable layout algorithm, Constrained Multi-Dimensional Scaling (CMDS). We conducted a user study that assessed the effectiveness of CorrelatedMultiples in visual search and comparison among multiple items, and quantitatively evaluated the quality and performance of CMDS relative to previous layout methods. We showed the benefits of CorrelatedMultiples for applications in stock market trend analysis, census demographics and climatology.

Chapter 4: The Effects of Representation and Juxtaposition on Graphical Perception of Matrix Visualization

4.1 Motivation

A network is an abstract data type that represents entities as nodes and their relationships as edges. Examples include social networks, computer networks, biological networks, and organizational networks. Network visualization has become an important research topic aiming to gain an effective overview of complex relational data [6, 27, 55, 56, 57]. Introduced by Jacques Bertin [13], adjacency matrices offer an interesting alternative to conventional node-link diagrams, which can suffer from visual cluttering due to node overlapping and edge crossing. An adjacency matrix shows how nodes are connected together through the intersection of the corresponding row and column. When the connections are undirected, the adjacency matrix is symmetric with respect to the main diagonal. Consequently, the same information is shown in both the upper and lower triangular matrices. Adjacency matrices have been shown more readable than node-link diagrams for many graphical-perception tasks, particularly when networks are dense [4, 40, 77]. Since many real-world networks are naturally dynamic and associated with multiple attributes, analyzing multiple networks at once is a common yet difficult task. Therefore, visualizations that aid users to compare and contrast multiple networks are of great importance.

Juxtaposition is an effective visual design that encourages side-by-side visual comparison of multiple facets of a complex data set, without overplotting or occlusion that may occur in *superimposition*, which overlays many objects in a single visualization. Although conventional side-by-side juxtaposition, or small multiples [13, 143], has been applied to adjacency matrices for comparative analysis [6], it predominantly relies on the use of the viewer's memory and attention shifts to make connections between repeated objects [42]. When too much of the comparative burden is added onto the viewer's mental effort, he or she can fail to detect changes even if the information is well represented in the visualization [125]. Designing appropriate juxtaposed visualization for adjacency matrices, which allows visual processing to connect patterns across multiple matrices with less mental effort, is still an open problem. Furthermore, an effective representation of multiple adjacency matrices is an instance of fundamental visualization research: making more effective use of display space to increase the amount of data with which users can effectively work [54, 143]. However, it remains unclear how different matrix representations and juxtaposition designs affect the ability of users to quickly and reliably understand the underlying information.

In this chapter[1], we evaluate the representation and juxtaposition designs for visualizing adjacency matrices through a series of controlled experiments. We investigate the effects of matrix representation on the speed and accuracy of performing graphical-perception tasks. Based on *human symmetric perception*, an automatic visual process that forms an integral part of perceptual organization [150], we propose two alternative juxtaposition designs to the conventional side-by-side juxtaposition, and study how users perform visual search and comparison tasks regarding these juxtaposition types. Our results show that triangular

[1]Major portions of this chapter were previously published in Liu et al. [92].

matrices are as effective as square matrices, and the juxtaposition types (*side-by-side*, *back-to-back*, and *complementary*) perform differently. With the design guidelines derived from our studies, we present a compact visualization termed *TileMatrix* for juxtaposing a large number of matrices, and demonstrate its effectiveness in analyzing multi-faceted, time-varying networks using real-world data.

4.2 Background

The node-link diagram and adjacency matrix are the most popular for network visualization. Node-link diagrams are generally suitable for sparse networks while adjacency matrices are more effective for dense networks, as shown by Ghoniem et al. [40], Keller et al. [77] and Alper et al. [4]. Many research efforts have recently been made to enhance the usability and readability of matrix-based network visualization. MatrixExplorer [56] couples a node-link diagram and an adjacency matrix representation of the same network by juxtaposed views. MatLink [57] augments matrix representation with links connecting nodes in lines and columns to reconcile the difficulty of path-related tasks in matrix visualization [40]. NodeTrix [55] visualizes community structures (dense sub-networks) as adjacency matrices, which are then reconnected with links that represent the sparse parts of the network. Dinkla et al. [27] exploited the structural characteristics of gene regulatory networks to design compressed adjacency matrices. Behrisch et al. [11] placed matrices of high similarity together in a projection space. However, such automated approach cannot reveal multiple facets of the data, and the order in time is lost after projection, which can make it difficult to identify trends of networks over time. Cubix [6] stacks adjacency matrices over time to form a 3D space-time cube, and supports interactive transitions to projected 2D juxtaposed views. While most previous works employed square matrices for

visualizing networks, we explored the effects of triangular matrices and their use in designing compact juxtaposed visualization for showing more readable networks within a limited display space.

4.3 Study 1: Evaluating Matrix Representation

As stated in the beginning, a symmetric adjacency matrix can be represented in either a square matrix (Figure 4.1 (left)) or a triangular matrix (Figure 4.1 (right)). The goal of the first experiment is to determine the impact of the matrix representation on the user's graphical perception: how does the choice of square or triangular matrices affect user performance?

4.3.1 Tasks

To keep the experiment manageable in time and effort for the participants, we did not include complex tasks related to finding paths and common neighbors since they can be difficult for the given matrix visualization [40]. Rather, such tasks can be accomplished more easily with node-link diagrams [40] or with additional visual cues [57]. Since we were interested in studying the effects of two different representations, tasks that are less relevant to the representation, such as counting the number of nodes or finding a node given its label [40], were also excluded in our experiment. Finally, three generic tasks were selected in an attempt to capture both overview and detail use cases of matrix visualization:

(T1) *How many communities can you identify in the matrix?*

(T2) *Which node has the maximum number of neighbors?*

(T3) *Are the two specified nodes connected in the matrix?*

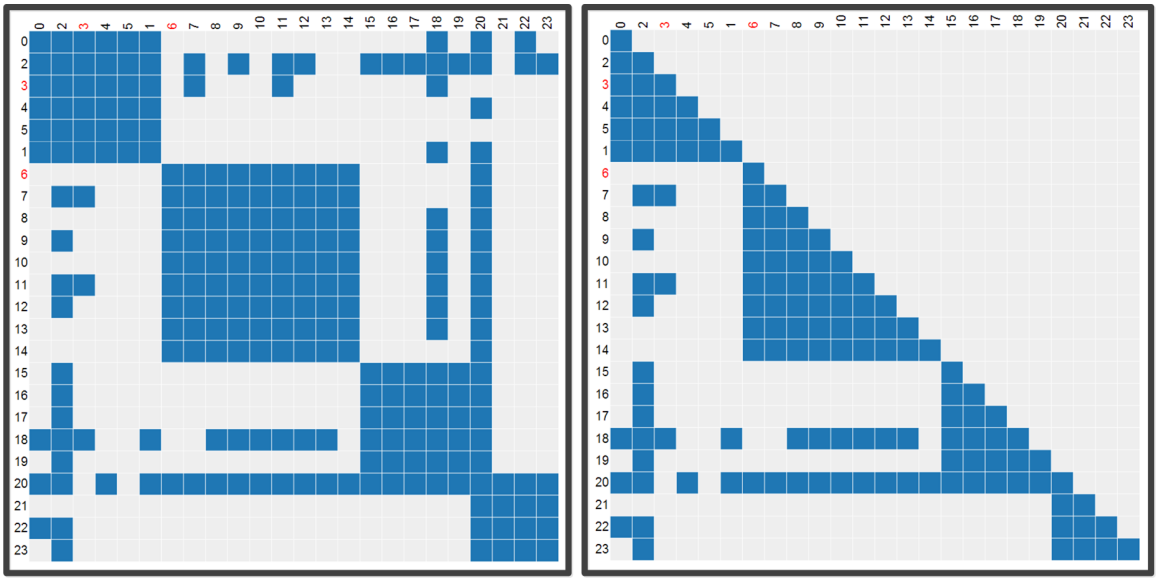


Figure 4.1: Adjacency matrix representations. Left: a square matrix representation. Right: a triangular matrix representation.

T1 is an overview graphical-perception task that represents a common use case of matrix visualization for community detection in networks [55]. T2 is an exploratory visual search task that requires browsing through the visualization to find information. T3 is a confirmatory visual search task that asks the users to make judgments once they locate the specified targets. T1 and T2 are related to the interpretation of blocks (communities) and lines (a node's neighbors), which are categorized as structural features by Mueller et al. [108]; T2 and T3 were also used in previous studies [40, 77].

4.3.2 Hypotheses

(H1) *The square matrix will perform as well as the triangular matrix for overview perception (T1).* For viewing the shapes of communities in a matrix, we believe that squares and triangles should be equally easy for perception.

(H2) *The square matrix will outperform the triangular matrix for exploratory search (T2).* For square matrices, users scan through rows or columns to view the neighbors of one node; for triangular matrices, users need to follow an L-shaped path to view the corresponding neighbors. We expect that following a straight line in rows or columns will have better user performance than following an L-shaped path.

(H3) *The square matrix will outperform the triangular matrix for confirmatory search (T3).* Square matrices encode one connection between two nodes twice in the visualization, while triangular matrices represent the connection exactly once. We predict that the repetition of such information will have a direct effect on user performance.

4.3.3 Experiment Design

The study was conducted as a within-subjects experiment with 2 experimental conditions (matrix representations) and 10 repetitions for each condition. For each repetition, the participant was presented with only one condition. We counter-balanced the selection of condition in the 10 repetitions so that each participant performed the same number of repetitions for both conditions while the choice of condition is random.

In the spirit of classic graphical perception experiments [35, 54, 66], we evaluated the different visual representations alone, disabling selecting, highlighting, zooming, and other interactive operations. Ghoniem et al. [40] have shown that in most cases, matrix visualization was insensitive to size and density variation and no interaction between size and density was found for the matrix representation. Hence, while the size and density of the matrices vary across repetitions for creating diverse experimental datasets, the effects due to size and density were not explicitly studied in our experiment.

We follow the convention of perception studies [54, 66, 86] in using synthetic data to allow control over the characteristics of experimental datasets. We first generated a random number of communities of varying sizes, then added a random number of edges for randomly selected nodes. In order to eliminate any ambiguity with respect to T2, which is to find the most connected node, we added an extra 20% of edges to the most connected node in every matrix. We labeled the nodes numerically according to the order of their creation, and ordered the nodes in such a way that those from the same community are placed together to preserve the community structures in the matrix. In this way, we obtained repetitions of varying number of nodes (from 20 to 50) and edges (density from 0.2 to 0.5), as well as varying number of communities (from 3 to 6). Specifically, the labels of the two specified nodes for T3 are highlighted in a distinct color. Figure 4.1 shows one example of the datasets we used for this experiment.

20 subjects (14 males, 6 females) were recruited for this experiment. The subjects were graduate students, aged 24 to 32 years. Half of the subjects have backgrounds in computer science, and most are using a computer more than 20 hours per week. All subjects have normal or corrected-to-normal vision. 40% of the subjects said they were familiar with matrix visualization.

We deployed the experiment on the web using HTML5 and JavaScript. Each subject used their own machine and browser, and the visualization images were scaled to fit the corresponding screen space before the test. Due to our within-subjects design, each user performed tasks regarding different representation types on the same screen he or she had. Hence, the influence of screen size has been alleviated when studying the effects of representation types. Prior to the experiment, subjects viewed a tutorial that gave a basic explanation of the two matrix representations, and they performed some training tasks to

Table 4.1: RM-ANOVA analysis of results for Study 1.

	Factor	$F_{1,19}$	p
Overall	Time	0.82	0.37
	Accuracy	0.16	0.69
T1	Time	0.08	0.79
	Accuracy	0.40	0.53
T2	Time	0.48	0.49
	Accuracy	0.29	0.60
T3	Time	1.56	0.22
	Accuracy	0.18	0.68

get familiar with the user interface of the experimental system. We asked the subjects to answer as quickly as possible while trying to make answers accurate. After the subjects finished tasks for all repetitions, they were asked to participate in a semi-structured interview.

4.3.4 Results and Discussion

We measured the time participants needed to complete each task and the correctness of their reported answers. User performance measures were evaluated using Repeated Measures Analysis of Variance (RM-ANOVA) to test for significant effects. For each participant's performance, we used the average of the repetitions for the following analysis. We report completion time and accuracy in Figure 4.2 and our RM-ANOVA analysis in Table 4.1.

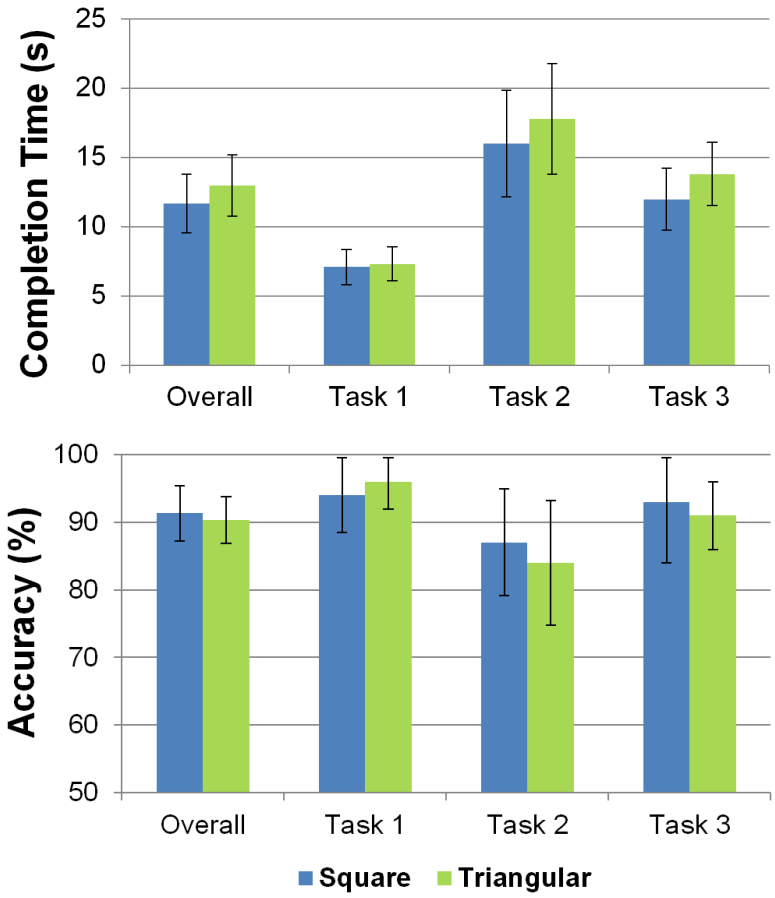


Figure 4.2: Mean completion time and accuracy for Study 1.

Our first hypothesis was that the square matrix would perform as well as the triangular matrix for overview perception. This hypothesis was confirmed. We found that the accuracy and completion time were almost the same across the two representations, which indicates that users were able to interpret communities in both representations.

One surprising outcome of our study is that the matrix representation type has no significant effect on the completion time of exploratory search tasks like finding the most connected node in a matrix. In other words, it seems that participants did not become

significantly slower when following the L-shaped path to view one node's neighbors. Participants commented that tracking neighbors for one node was a little bit hard at start, but their speed got faster after the training trials.

We also hypothesized that the square matrix would outperform the triangular matrix for confirmatory search. Contrary to our hypothesis, our quantitative results did not show significant differences between the matrix representations. Rather, these results indicate that the participants performed equally well in the confirmatory search, regardless of the matrix representations.

Overall, results from our first experiment revealed that users were able to interpret triangular matrices correctly, and their task performance with the triangular matrix was comparable to that with the square matrix.

4.4 Study 2: Evaluating Matrix Juxtaposition

The goal of the second experiment is to investigate the effects of juxtaposition types in the presence of multiple adjacency matrices. More specifically, we are interested to see how users perform visual search and comparison tasks under different juxtapositions. Is there a benefit to introducing more types of juxtapositions other than conventional side-by-side juxtaposition?

In our previous experiment we found that square and triangular representations of adjacency matrix had comparable accuracy and completion time, so we removed square matrices from consideration in this experiment and focused on comparing triangular matrices of different juxtaposition types.

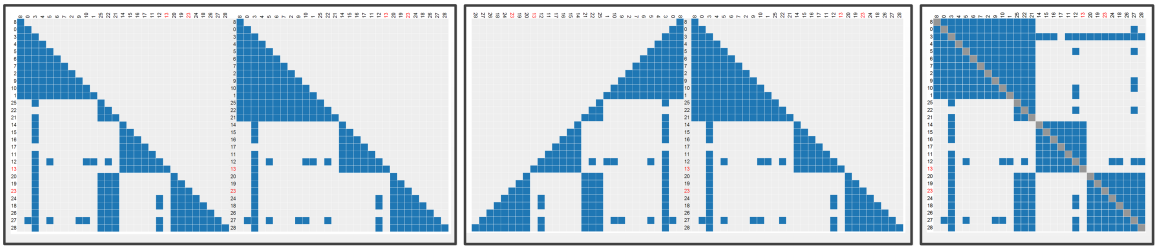


Figure 4.3: Adjacency matrix juxtapositions. Left: side-by-side juxtaposition (SID). Middle: back-to-back juxtaposition (BAC). Right: complementary juxtaposition (COM).

4.4.1 Matrix Juxtaposition

We divide the design space of comparative juxtaposition into three general types, based on how the relationships between the related parts of different adjacency matrices are encoded:

Side-by-side juxtaposition (SID), also known as small multiples [143], repeats the same representation multiple times without any modification to the design (Figure 4.3 (left)), and has been applied to various domains such as system management, quality control, and medical record analysis [143]. This *translational juxtaposition* is the conventional design of comparative juxtaposition.

Back-to-back juxtaposition (BAC), which reverts the order of the rows or columns in one of two matrices to form a symmetric composition (Figure 4.3 (middle)). This design is motivated by human symmetric perception, which is an automatic visual process that forms an integral part of perceptual organization [150]. Such *symmetric juxtaposition* has been shown effective in identifying similarity and contrasts in computer-aided diagnosis. For example, radiologists are using the differences between the symmetric juxtaposition of left

and right breasts in mammograms (photographs of breasts made by X-rays) to help detect certain malignant breast cancers [132].

Complementary juxtaposition (COM), which takes two triangular matrices together to form a compact square matrix. Consequently, the rows and columns in the two matrices are complementary, and the positions of cells in the two matrices are symmetric with respect to the main diagonal (Figure 4.3 (right)). In other words, two triangular matrices form an asymmetric square matrix. This design is a variant of back-to-back juxtaposition, and only applies to triangular matrices.

We think these three juxtaposition types are *fundamental*: they define the juxtaposed relationship of two matrices (translational juxtaposition for **SID** or symmetric juxtaposition for **BAC** & **COM**), representing the basic cases for comparative visualization. More importantly, they provide the building blocks that can assemble complex juxtaposition of many more matrices. In other words, the three juxtaposition types can be combined to create hybrid juxtaposed visualization.

4.4.2 Tasks

Typical visual comparison tasks often require users to scan and compare related elements simultaneously. Three tasks, modified from the tasks selected in our previous experiment, were included in this experiment:

(T4) *Does the largest community have the same number of nodes in the following matrices?*

(T5) *Does the most connected node have the same number of neighbors in the following matrices?*

(T6) *How many times (0, 1, or 2) are the two specified nodes connected in the following matrices?*

In each task, users were asked to identify multiple targets within two matrices presented in one of the three juxtaposition types, and determine the similarity or difference of the targets. As in our prior experiment, these tasks were selected to represent overview and detail use cases of matrix visualizations, with a balance between task complexity and suitability.

4.4.3 Hypotheses

(H4) *For tasks involving comparison of structures and patterns in matrices (T4, T5), BAC & COM (symmetric juxtaposition) will outperform SID (translational juxtaposition) in task completion time.* The strength of symmetric perception is that it allows easier comparison of shapes across objects. We believe that the mental alignment due to symmetric perception will result in better completion time by connecting patterns and reveal their differences across matrices.

(H5) *When searching and comparing a specific target in matrices (T6), SID, BAC and COM will have comparable performance.* Comparing a specific target requires accurately locating the same item across matrices. We predict that symmetric juxtaposition requires more mental effort to locate a specific target across matrices. However, we think that the easier mental alignment due to symmetric perception compensates this additional effort, resulting in a performance comparable to that obtained using translational juxtaposition.

4.4.4 Experiment Design

The experiment used a 3 (juxtaposition) $\times$ 3 (task) within-subjects design with 9 repetitions. For each repetition, the participant was presented with only one condition. We

counter-balanced the repetitions so that each participants performed the same number of conditions for three conditions while the selection of condition in each repetition was random.

We followed the design choices we made in our previous experiment. We first generated matrices that have varying numbers of nodes (from 20 to 50) and edges (density from 0.2 to 0.5), and varying numbers of communities (from 3 to 6). Then for each generated matrix, we randomly assigned several nodes to a different community and added edges for randomly selected nodes to get a new matrix. The two matrices were then grouped using one of the three juxtaposition types. Figure 4.3 shows one example of the datasets we obtained for this experiment.

We recruited 28 subjects (17 males, 11 females, aged 24 to 39 years), who are graduate students or professionals from the fields of computer science, electric engineering, chemistry, statistics, geographic information science, to name a few. Half of the subjects had previously participated in the first experiment. 36% of the subjects said they had experience with multiple matrix visualization. The experiment follows a similar procedure to the prior experiment. After the subjects finished all tasks, they were asked to rate their satisfaction with **SID**, **BAC** and **COM** in each task on a questionnaire (e.g., *Which view do you think is the best to identify the size difference of the largest community? (A: side-by-side; B: back-to-back; C: complementary; D: no preference)*). Finally, they were asked to participate in a semi-structured interview.

4.4.5 Results and Discussion

We report completion time and accuracy in Figure 4.4. Statistical significance was tested for all groups and each pair respectively. Table 4.2 summarizes all results and highlights the significant ones. The qualitative feedback is shown afterwards in Table 4.3.

In H4 we hypothesized that **BAC** & **COM** would be faster than **SID** due to their symmetric alignment for detecting changes of structures and patterns in matrices. This hypothesis was confirmed. We found that **SID** took significantly longer time than **BAC** & **COM**. This result validates our design goals: facilitating identification of the repeated patterns and differences of connectivity between juxtaposed matrices. Subject ratings (Table 4.3) confirmed that most participants preferred either **BAC** or **COM** in accomplishing T4 and T5. Participants commented that **BAC** and **COM** are helpful in connecting the left and right matrices, while **SID** requires more mental effort in certain cases.

Our results confirmed H5: **BAC** & **COM** would not have a negative impact on performance when searching and comparing a specific target in matrices. We found no significant difference between the juxtaposition types in task completion time or accuracy. Participants do not seem to have a strong preference in choosing a particular juxtaposition for T6. Accuracy seems to stabilize at a high rate for each task in any juxtaposition. It indicates that the participants were equally careful in visual comparison, regardless of juxtaposition type.

4.5 Design Implications

Based on our experimental results, we offer the following implications for designing adjacency matrix visualizations:

Triangular representation does not hamper graphical perception of adjacency matrices. One unexpected result was that the triangular matrix neither slowed down the task

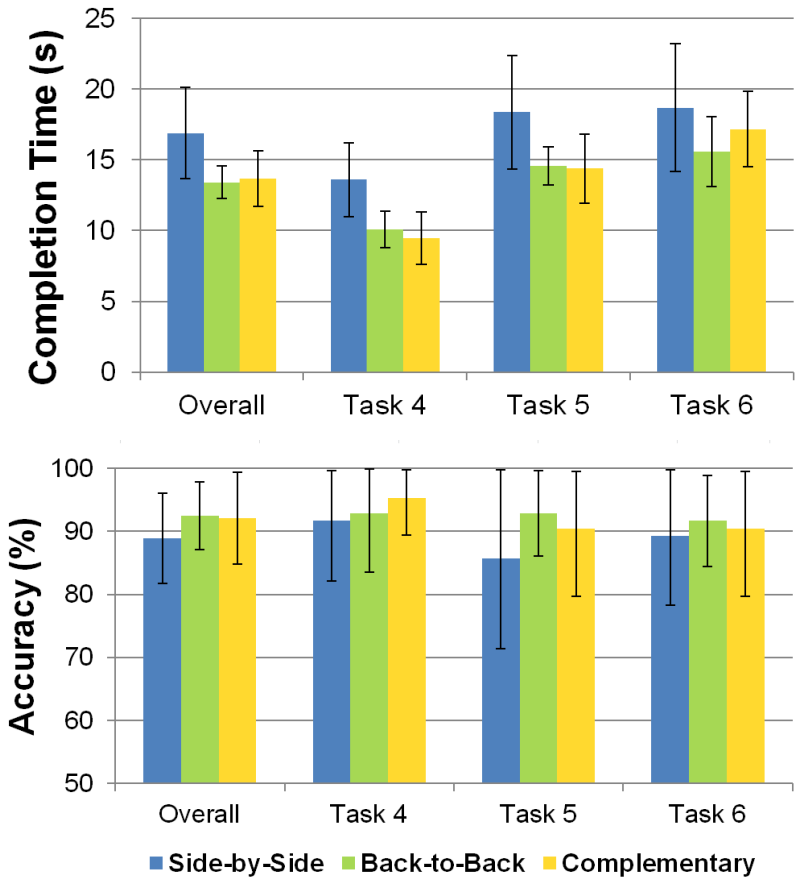


Figure 4.4: Mean completion time and accuracy for Study 2.

completion time nor hurt the accuracy. The triangular representation cuts the size of the matrix in half without any observed downside, as long as the viewer learns how to interpret the triangular matrix.

Symmetric juxtaposition rather than translational juxtaposition should be preferred for detecting changes of structures and patterns. The mental alignment due to symmetric perception is able to connect patterns and structures across matrices, which facilitates visual comparison with less mental effort.

Table 4.2: RM-ANOVA analysis for Study 2 (s:SID, b:BAC, c:COM).

	Factor	$F_{2,54}$	$p_{s,b,c}$	$p_{s,b}$	$p_{s,c}$	$p_{b,c}$
Overall	Time	4.81	**0.01**	**0.01**	**0.03**	0.77
	Accuracy	0.59	0.56	0.30	0.42	0.91
T4	Time	8.47	**0.001**	**0.003**	**0.001**	0.49
	Accuracy	0.32	0.73	0.82	0.41	0.57
T5	Time	4.30	**0.02**	**0.02**	**0.03**	0.86
	Accuracy	0.73	0.49	0.25	0.49	0.63
T6	Time	1.46	0.24	0.12	0.46	0.26
	Accuracy	0.10	0.91	0.64	0.84	0.81

Table 4.3: Subjective user preference for tasks in Study 2.

	SID	BAC	COM	None
T4	0%	46%	50%	4%
T5	10%	36%	40%	14%
T6	21%	21%	18%	40%

Complementary juxtaposition is beneficial for optimizing utilization of display space. Since complementary juxtaposition doubles the data density (data marks per display area) compared to other juxtaposition types, we advocate its use when space constraints warrant.

4.6 TileMatrix: Creating Compact Visualization by Tiling the Matrices

With the design implications, we devise a compact visualization — *TileMatrix*, coupling the side-by-side juxtaposition, back-to-back juxtaposition and complementary juxtaposition to display a large number of adjacency matrices. The TileMatrix representation is inspired by the physical act of laying tiles to cover floors, walls, ceilings and roofs. A *tile* is generally designed in an interlocking pattern so that multiple tiles fit together to have an aesthetic appearance. In TileMatrix, one tile is a triangular matrix. Every two adjacent matrices are placed in either back-to-back juxtaposition or complementary juxtaposition, resulting in a hybrid juxtaposition that also includes side-by-side juxtaposition . In the following sections, we first discuss typical usage scenarios for displaying a large number of matrices, and then describe our TileMatrix system. We demonstrate its effectiveness in a case study using real-world data, and report initial user feedback.

4.6.1 Usage Scenarios

A weighted network associates a weight with every edge in the network, representing the strength of the connection between the entities [4]. In many network applications, the weights of edges can be quite complex and dynamic. On the one hand, the weight of the connection between the entities can change as multiple facets are associated with the connection. For instance, consider a collection of National Basketball Association (NBA) players as the entities, who are connected either strongly or weakly based on the similarity of their performance. Each player is associated with multiple performance facets: points, rebounds, assists, steals, blocks, and many others. Therefore, players can have different weights of connections on different facets. On the other hand, the weight of a connection

between the entities can change over time. This is also seen in the NBA players example since their performance often vary in different years.

Visual analysis of such multi-faceted, time-varying weighted networks is challenging as the number of networks grows in both data and temporal domains. Although we can always reduce the number of facets using dimension reduction techniques or aggregate networks over time, we learn little about the heterogeneity nor the dynamics of the network. The reduced dimensions do not always have a semantic interpretation, and the temporal differences are hidden in temporal aggregation. Consequently, we understand neither how to interpret the relationships from multiple perspectives nor how certain temporal trends are formed.

TileMatrix offers one alternative solution to visualizing multi-faceted, time-varying weighted networks without losing information — the matrices of networks are tiled simultaneously in two directions of the display space: matrices in different facets are tiled in columns while matrices over time are tiled in rows. In this way, the viewer can examine and compare networks of multiple facets at a particular time step (horizontally), as well as networks of a particular facets over time (vertically). The repeated patterns and differences in the relationship of entities can be identified, in both data and temporal domains.

4.6.2 Case Study

To demonstrate the effectiveness of TileMatrix, we used our system to explore the NBA statistics dataset [141], which consists of 16 performance facets for NBA players: GP (Games Played), MIN (Minutes Played), PT (Points Scored), AST (Assists), REB (Rebounds), STL (Steals), BLK (Blocked Shots), TO (Turnovers), ORE (Offensive Rebounds), DRE (Defensive Rebounds), FGA (Field Goals Attempted), FGM (Field Goals Made), FTA

(Free Throws Attempted), FTM (Free Throws Made), TPA (Three Point Attempted), and TPM (Three Point Made). To view a relatively large number of matrices with TileMatrix, in this study we extracted players that have relatively longer NBA careers (at least 15 years compared to an average career length of 5 years) from 1989 to 2003. This gave us 11 NBA players: Horace Grant, Reggie Miller, Vlade Divac, Avery Johnson, Glen Rice, Karl Malone, Scottie Pippen, Charles Oakley, Kevin Willis, Rod Strickland, and Mark Jackson. The similarity networks for all pairs of players were computed regarding each facet for each year. The corresponding TileMatrix contains $16 * 15 * 11 * 11/2 = 14520$ cells. The color of self-connections is set to grey. Matrices of adjacent facets are assigned similar colors while distant facets are colored differently for overall appearance.

By analyzing the semantic meaning of the performance facets, we roughly categorized them into three groups based on the criteria of the performance measurement: *activity* = {GP, MIN}, *score* = {PT, FGA, FGM, FTA, FTM, TPA, TPM}, *aid* = {AST, STL, BLK, REB, ORE, DRE, TO}. Therefore, we reordered the facets in TileMatrix to place related facets in the same category closer. To highlight the contrasts between highly and weakly connected nodes, we reordered the nodes by the weighted average degree.

Figure 4.5 shows the resulting TileMatrix visualization after reordering. Facets are labeled on the top of TileMatrix while years are on the left. The names of the players are also labeled accordingly. From the TileMatrix visualization, we had a few key findings on the multi-faceted, time-varying relationships of the NBA players that had relatively longer careers:

Temporal trends regarding a particular facet. To understand the time-varying similarity/dissimilarity of players' performance with respect to one facet, we can simply view matrices form a column of TileMatrix. For example, the left-most column of TileMatrix in

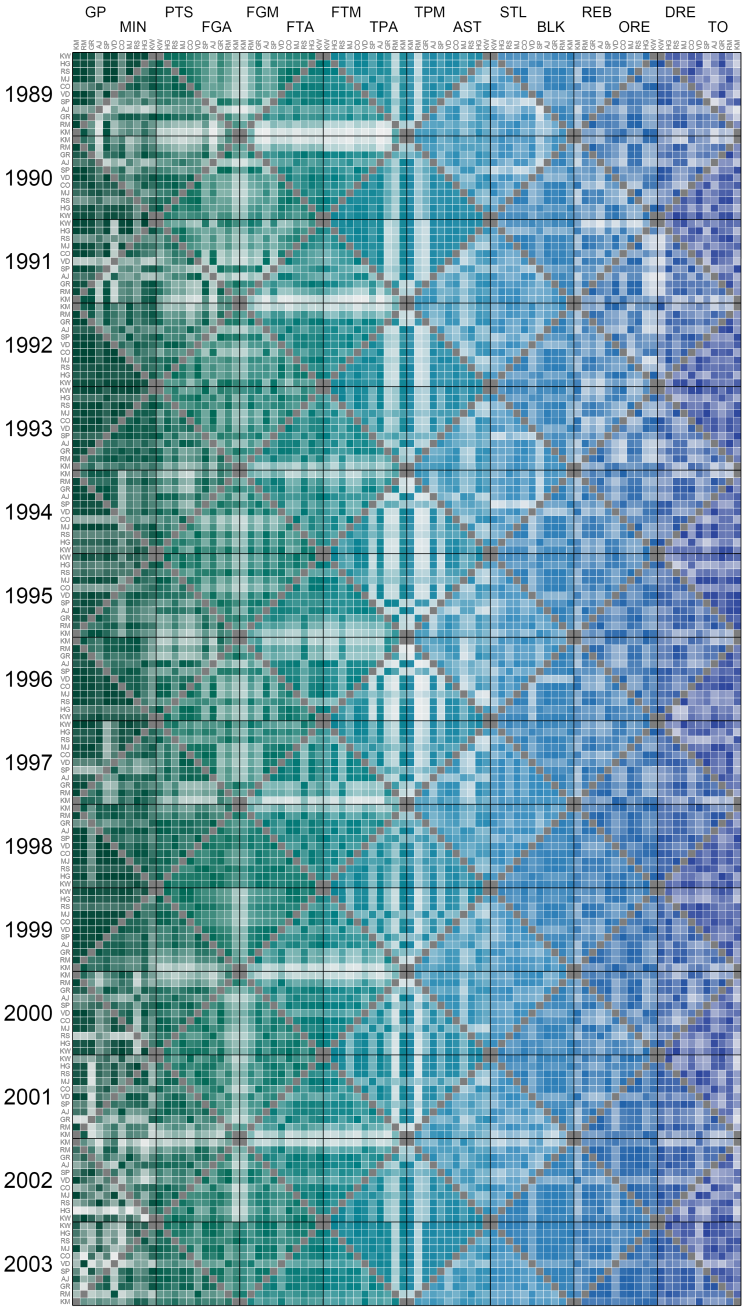


Figure 4.5: The TileMatrix visualization of 16-faceted similarity networks of National Basketball Association (NBA) players from 1989 to 2003. Matrices of different facets are tiled in columns, while matrices of different years are tiled in rows. The color opacity encodes the strength of the connection.

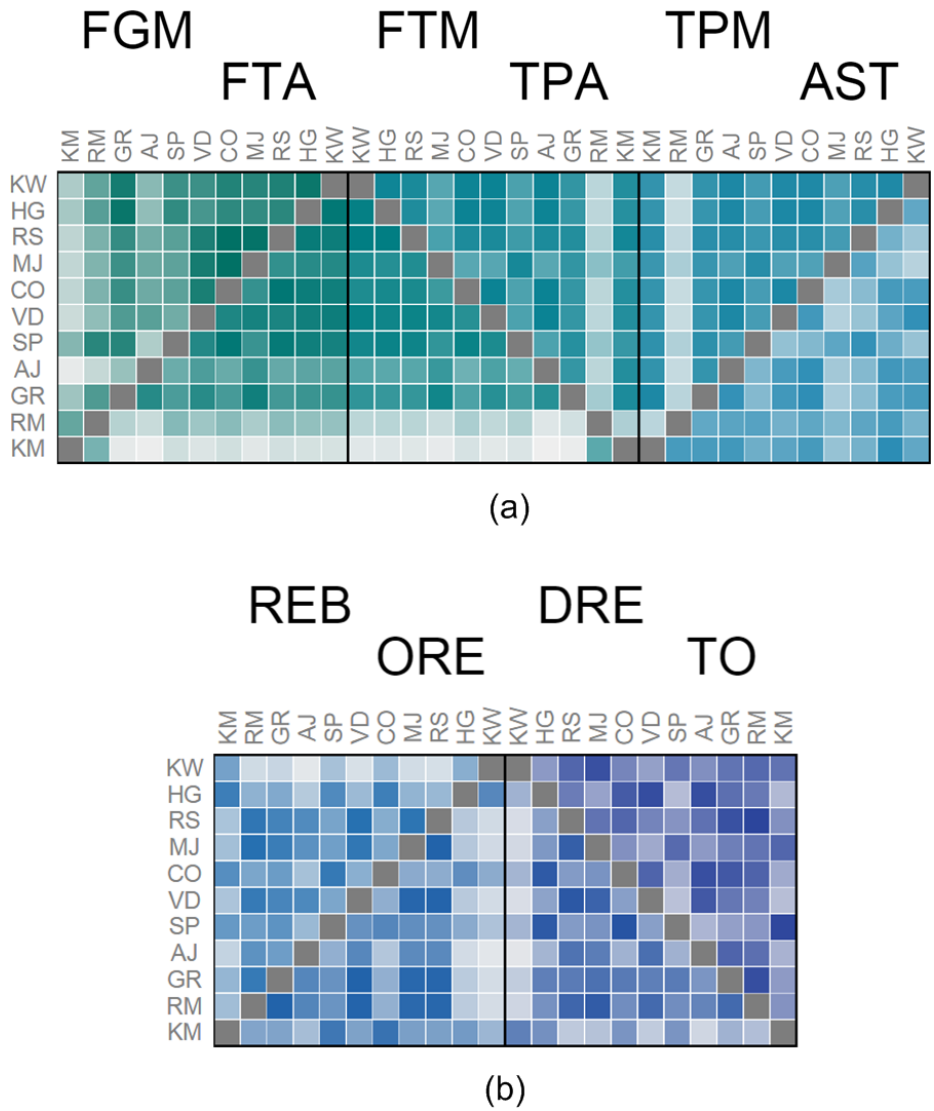


Figure 4.6: Two zoom-in views of the TileMatrix visualization in Figure 4.5. (a) Matrices of 6 selected facets (FGM, FTA, FTM, TPA, TPM, AST) in 1989. (b) Matrices of 4 selected facets (REB, ORE, DRE, TO) in 1991.

Figure 4.5 shows the time-varying similarity networks based on the number of games the players had played, from which we can observe the following trend: (a) in 1989, 1991, and 1997: one or two players had played quite a different number of games compared to the rest of players, as the their connections to the others were relatively weak (e.g., CO (Charles

Oakley) and AJ (Avery Johnson) in 1989, VD (Vlade Divac) in 1991, and SP (Scottie Pippen) in 1997); (b) in 1992, 1993, 1996 and 1999: the number of games played were almost the same among the players, since the matrix cells show very strong connections; (c) during 2000-2003: players started to have distinct degrees of activity, and the differences seemed to reach a peak at the end of this time window (e.g., in 2003, most players were strongly connected to only a few other players while weekly connected to the majority). This could be partly because some players had played for many years, and their performance became diverse compared with their early career.

Repeated patterns across multiple facets at a specific time. As implied from our experiments, the symmetric juxtaposition has significant effects on connecting patterns and structures across matrices. To view the multi-faceted similarity/dissimilarity of players' performance at a specific time, we can focus on one row of matrices in TileMatrix. For instance, the first row of TileMatrix in Figure 4.5 shows the multi-faceted similarity networks of players in 1989. Approximately repeated visual patterns can be identified between the following groups of facets respectively: FGA & FGM, FTA & FTM, and TPA & TPM. From a zoom-in view of the facets in Figure 4.6 (a), we learned that: if one player (e.g., RM (Reggie Miller)) behaved differently from the others in one facet (e.g., TPA), he was very likely to perform differently from the others in another related facet (e.g., TPM); if the performance of a group of players (e.g., KW (Kevin Willis), HG (Horace Grant), RS (Rod Strickland), MJ (Mark Jackson), CO (Charles Oakley), VD (Vlade Divac)) were similar to each other in one facet (e.g., FTM), such similarity relationship could also be found in another related facet (e.g., FTA).

Contrasts between multiple facets at a specific time. Another implication from our experiments is that the symmetric juxtaposition also helps reveal contrasts between adjacent facets. Take the similarity networks on ORE and DRE in 1991 as an example (shown in Figure 4.6 (b)). We observed that the degree of dissimilarity to other players are different between offensive rebounds and defensive rebounds for HG (Horace Grant), while it roughly stayed the same for KW (Kevin Willis). A comparison between REB and TO for the connections of KW (Kevin Willis) also highlights the contrasts between these two facets: his performance was more different from others on rebounds than on turnovers.

Temporal trends regarding multiple facets. Because of the two dimensional tiling layout in TileMatrix, we are able to examine the temporal trends of adjacent facets, by viewing matrices in adjacent columns as a whole. For example, the columns of TPA and TPM of TileMatrix in Figure 4.5 shows a trend for the dissimilarity of players' performance regarding these two facets: (a) starting from 1989, only one player RM (Reggie Miller) performed differently from the others; (b) the difference among players became more and more obvious over the years until it reached top in 1996, when three players performed very differently from the others regarding both three point shots and scores; (c) after that, the degree of dissimilarity among the players dropped; (d) finally in 2003, RM (Reggie Miller) once again became the only one that performed differently compared with the other players, just like how the trend started in 1989. Comparisons between other related facets also revealed interesting temporal trends.

4.6.3 Informal User Feedback

We observed six users (5 males, 1 female) viewing the TileMatrix in Figure 4.5, and a conventional side-by-side juxtaposed view of square matrices (denoted as SidMatrix)

in Figure 4.7. SidMatrix displays the same number of matrices as TileMatrix with the same color encoding. We explained the designs of TileMatrix and SidMatrix to the users, and how the information of the multi-faceted, time-varying similarity networks of NBA players is presented. Our task was very informal, simply asking them to comment on their understanding and visual comfort of the two visualizations, and how easily they could find interesting trends in each visualization. The two visualizations were uniformly scaled to the same width and put side by side for direct comparison.

As initial feedback, all six participants were able to understand TileMatrix as well as SidMatrix. One general feedback is that a cell in SidMatrix is much smaller and less legible than that in TileMatrix. This is expected since the two visualizations display the same number of matrices, square matrices in SidMatrix take double space compared with triangular matrices in TileMatrix, and thus each cell in SidMatirx is displayed in half size of that in TileMatrix. Participants commented that TileMatrix was easier to see details. They were more comfortable in seeing larger cells with more effective information in TileMatrix. Participants also explained that with TileMatrix, it was easier to compare adjacent matrices when they are in back-to-back or complementary juxtaposition, since the symmetric layout helped them to see the subtle differences, which were not as obvious in SidMatrix. However, some participants reported that TileMatrix required more training effort to follow the L-shaped path when viewing the connections of a particular player, while they were used to following straight lines in SidMatrix. This is consistent with the findings in our first controlled experiment.

Our observations partly explained the tradeoffs of introducing hybrid juxtapositions of triangular matrices in visualizing a large number of networks. While such hybrid design makes better use of the display space, and may be useful for identifying repeated patterns

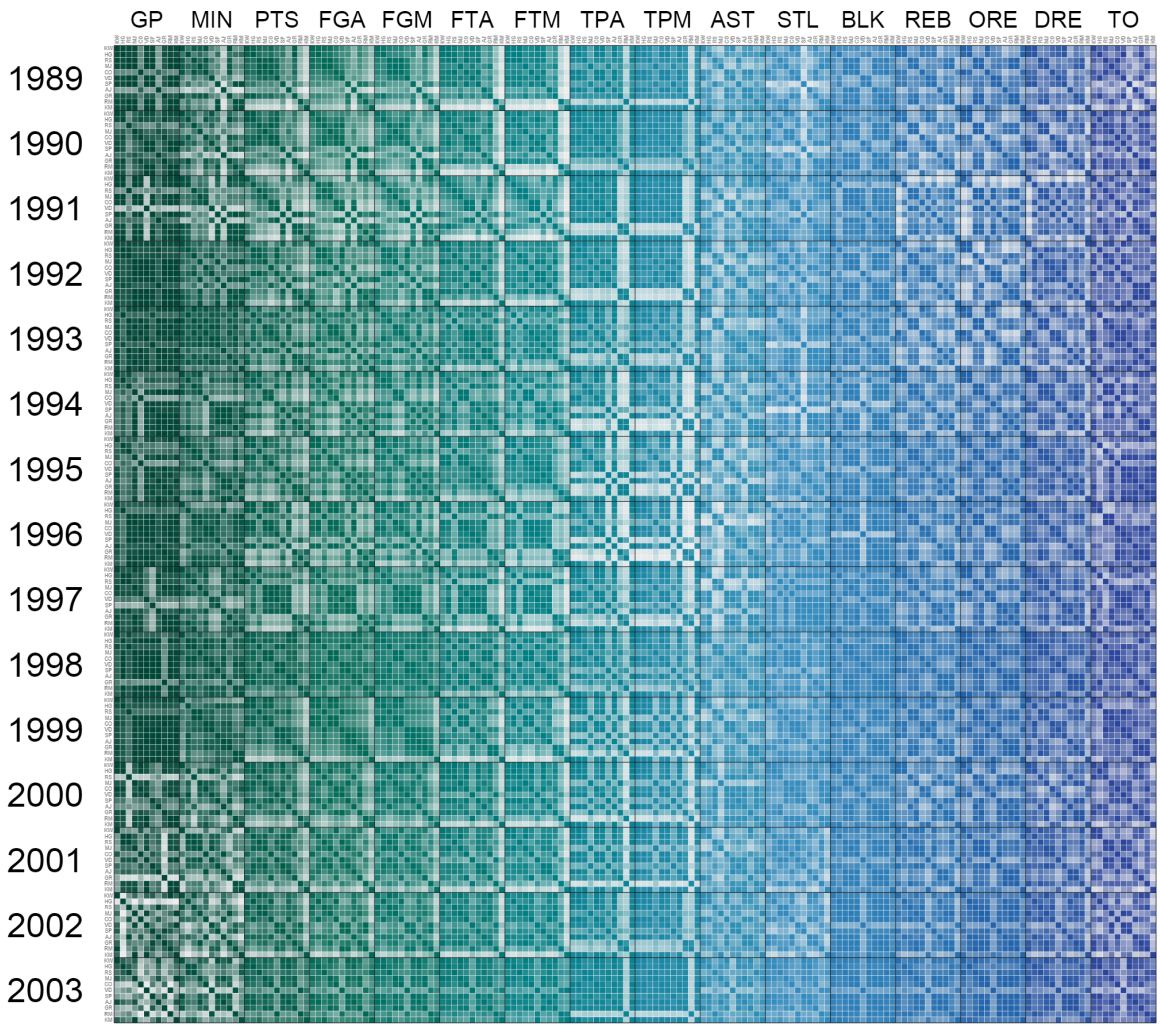


Figure 4.7: The side-by-side juxtaposed matrix visualization (SidMatrix) for the same data as TileMatrix in Figure 4.5.

and differences across multiple matrices more comfortably and clearly, it requires more training effort than the conventional side-by-side juxtaposed visualization.

4.7 Discussion

As stated in the beginning, the triangular matrix works when the connections in a network are undirected. As for directed networks, it is not possible to increase the data density by juxtaposing triangular matrices. Our experiments considered a small set of generic tasks for matrix visualization. We believe these findings generalize to a wider range of situations, but have not confirmed this empirically. Also we did not include interactive operations in our experiments but only focused on evaluating the effects of different representations and juxtapositions alone. Still, future work is needed to determine the additional effects when visualization is augmented by interaction.

Our TileMatrix design is effective in viewing the similarity and differences in matrices across facets and time, in particular for comparing matrices placed at nearby locations. On the other hand, using TileMatrix requires training to follow L-shaped paths when viewing one node's neighbors, and thus perhaps requiring more effort when tracing one node's neighbors across multiple matrices. Providing additional interactions such as highlighting one's neighbors across matrices when selecting a node could be helpful. Although the effect of colors used in TileMarix was not explicitly studied in our experiments, the design of TileMatrix can be easily extended, and the task performance with additional factors such as colors will be investigated in the future.

The scalability of TileMatrix depends on several factors: the number of attributes (M), the number of time steps (T), and the number of entities per attribute per time step (N). The total number of cells in TileMatrix is M * T * N * N / 2. Given a display screen, a tradeoff between N, M and T must be made when using TileMatrix, for example: (1) the network is small so many attributes and/or time steps are shown (as in our case study); (2) the network is medium sized so a medium number of attributes and time steps are shown; and (3) the

network is large, thus a few selected attributes and time steps are shown. We note that if the display screen cannot hold all cells, panning and zooming could be provided to view sub-regions on demand. Although we demonstrate the effectiveness of our TileMatrix design with a case study and informal user feedback, the trade-off between design complexity and task performance needs to be better understood in the future.

4.8 Summary

In this dissertation, we conducted two controlled experiments to assess the performance of adjacency matrices in two representations — square matrices and triangular matrices, and three juxtaposition designs — side-by-side juxtaposition, back-to-back juxtaposition, and complementary juxtaposition. We quantitatively measured speed and accuracy based on generic tasks in each experiment. The results showed that triangular matrices were as effective as square matrices, and the three juxtaposition types performed differently. We showed that back-to-back juxtaposition and complementary juxtaposition are generally a good choice for detecting changes of structures and patterns across matrices due to the mental alignment of symmetric perception. Based on the design guidelines derived from our studies, we propose a compact visualization termed TileMatrix for juxtaposing a large number of matrices, and show its benefits in analyzing multi-faceted, time-varying networks using real-world data.

Chapter 5: Association Analysis for Visual Exploration of Multivariate Scientific Data Sets

5.1 Motivation

Scientific simulations often create multiple variables describing different physical properties within the same spatial domain. Usually, certain scalar values from a subset of variables carry a greater importance to the understanding of the underlying phenomena than others [73]. For instance, in climate simulation, a hurricane typically forms as the *warm*, *moist* air over the ocean rises in the *low* air pressure area, which creates *strong* winds and *heavy* rainfall. However, without sufficient prior knowledge, it is generally difficult for users to select informative scalar values in a multivariate data space and understand how they are associated with scalar values of the other variables [64]. This makes visual exploration of multivariate scientific data sets a difficult task using only the existing visualization techniques such as direct volume rendering [88] and isosurface visualization [96]. More importantly, the heterogeneity and complexity of multivariate characteristics poses a unique challenge to this non-trivial problem, as it requires investigating the usually hidden associations between different variables and specific scalar values to understand the multi-faceted properties of the data sets.

Considering that each scalar value of a variable, or scalar in short, has a certain amount of information, encoded by the corresponding spatial data points, scalar-level interaction

can be viewed as the amount of shared information that flows from one scalar to another scalar of a different variable. In this sense, the roles that scalars play in the entire data set are not equally informative. An informative scalar should reveal information about the associated scalars of other variables and how they interact [14]. The directional aspect of information flows should be taken into account to determine how informative one scalar is compared with the other in an association. Commonly-used correlation measures such as correlation coefficients and mutual information mostly focus on studying the symmetric and average relationship between variables, and thus are unable to reflect the directional scalar-level relationships. Consequently, even for two correlated variables, it is still unclear how a specific scalar in one variable is associated with another scalar of the other variable. In addition, the enormous multivariate space complicates the search of potentially interesting scalars. Due to the large number of possible associations with scalars of many different variables, it is non-trivial to determine the relative informativeness that a scalar has in the entire multivariate domain.

In this chapter[1], we present a novel association analysis method that identifies informative scalars in multivariate data sets. To understand how the scalars interact with each other, we model the directional interactions between scalars of different variables as information flows based on *association rules*. Association rules are originally proposed by Rakesh Agrawal et al. [2] to analyze relationships of sale products from supermarket basket transactions, and confident rules can infer the sales of other products by observing the sales of a given product. In our method, we treat each scalar as a sale product and a co-occurrence of scalars between different variables as a transaction. We define the amount of information that flows from one scalar to another as the confidence of an association rule,

[1]Major portions of this chapter were previously published in Liu et al. [93].

and estimate it using conditional probability. We model all possible associations between scalars of different variables using a probabilistic association graph called PAGraph, and determine the informativeness of scalars in the PAGraph based on information propagation. The PAGraph is analogous to a social network in terms of two important factors — influence/informativeness and passivity/uniqueness: (1) in a social network, an individual of high influence can attract his audience and lead the interactions, while an individual of high passivity is often inert in response to the interactions; (2) in a PAGraph, an informative scalar can reliably infer the existence of other scalars of different variables, while a unique scalar can hardly be predicted by other scalars of different variables. To quantify informativeness and uniqueness from scalar-level associations in the PAGraph, we propose the Multi-Scalar Informativeness-Uniqueness (MSIU) algorithm based on an information propagation model in social networks [120]. Integrating our association analysis method with interactive visualization techniques, we present an exploration framework with multiple interactive views to explore the scalars of interest with confident associations in the multivariate spatial domain, and provide guidelines for visual exploration using our framework. We demonstrate the effectiveness and usefulness of our approach in exploratory analysis using three representative multivariate data sets.

5.2 System Overview

The main objective of this work is to guide visual exploration of multivariate data sets based on scalar-level associations. Essentially, we analyze scalar-level associations to determine the informativeness and uniqueness of scalars, which describe the information flows between scalars of different variables and reveal how they interact with each other in the multivariate domain. Based on scalar-level associations, we design multiple interactive

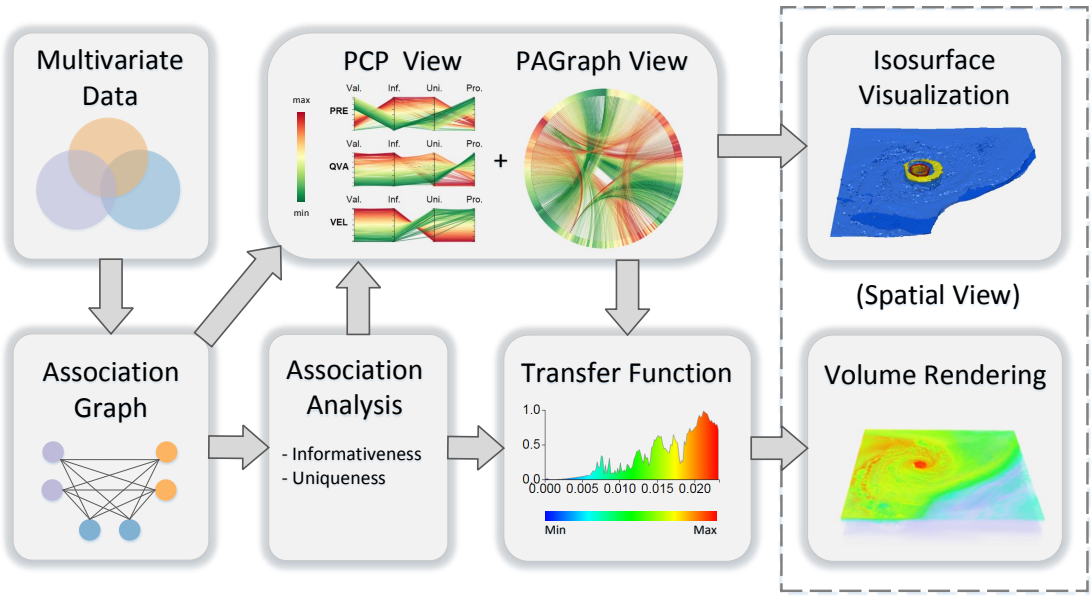


Figure 5.1: Overview of the analytical workflow. From the input multivariate data set, we model scalar-level associations using a probabilistic association graph, and apply the proposed association analysis method to quantify the informativeness and uniqueness of scalars. Multiple interactive views and spatial visualizations are created to explore scalar-level associations in the multivariate data set.

views to develop a framework for exploring scalar-level associations in multivariate data sets. Figure 5.1 illustrates our analytical workflow. From the input multivariate data set, we model all possible associations using a probabilistic association graph (Section 5.3.1 and 5.3.2). The informativeness and uniqueness of scalars are quantified by the proposed association analysis method (Section 5.3.3). To explore the scalars with associations of interest, the association graph with informativeness and uniqueness is used to create the PCP view and the PAGraph view, which are linked with isosurface visualization and direct volume rendering to reveal the spatial relationships of the selected scalars (Section 5.4.1).

5.3 Association Analysis in Multivariate Data Sets

In this section, we first model the directional interactions between scalars of different variables as information flows based on association rules, and describe how to represent scalar-level associations using a probabilistic multipartite graph. Then we introduce the concepts of informativeness and uniqueness to describe how information flows between scalars of different variables and how they are associated with each other in the multivariate domain. Finally we describe the Multi-Scalar Informativeness-Uniqueness (MSIU) algorithm to evaluate the informativeness and uniqueness of scalars.

5.3.1 Modeling Directional Interactions as Information Flows

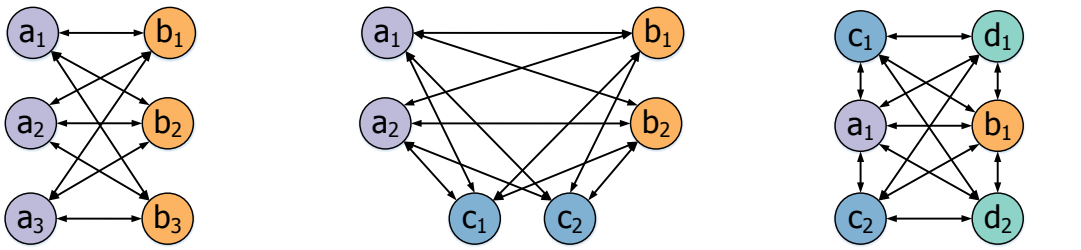
Considering that each scalar has a certain amount of information, encoded by the corresponding spatial data points, scalar-level interaction can be viewed as the amount of shared information that flows from one scalar to another scalar of a different variable. To quantify the information flows between scalars of different variables, we make use of *association rules*, which have been widely applied to discover interesting relationships among sale products in large transaction data sets [2]. An association rule is defined as an implication of the form $x \rightarrow y$. An example from the supermarket domain could be $Bread \rightarrow PeanutButter$, suggesting that a strong association exists between the sale of bread and peanut butter because many customers who buy bread also buy peanut butter in one transaction. For a given association rule $x \rightarrow y$, the strength of the rule can be measured in terms of its *confidence*, which can be interpreted as an estimate of the conditional probability $p(y|x)$. The confidence of an association rule can be used to infer the *dependency* that y has on x. The higher the confidence, the more likely it is for x to infer the existence of y. In other words, y is more likely to depend on x.

To make the analogy from the supermarket domain to the multivariate domain, let us treat each scalar as a sale product, and a co-occurrence of the scalars between different variables as one transaction. Considering two scalars x_i and x_j from two different variables X_i and X_j, two possible association rules $x_i \rightarrow x_j$ and $x_j \rightarrow x_i$ can be learned based on their occurrence in the multivariate space. The joint probability $p(x_i, x_j)$ reflects the amount of information shared between the two scalars, and the conditional probability $p(x_j|x_i)$ reflects the amount of shared information that flows from x_i to x_j. Naturally, $p(x_j|x_i)$ quantifies the confidence of an association rule $x_i \rightarrow x_j$ in the multivariate scenario. Unlike most traditional correlation analysis that only studies the average symmetric relationships between variables, such an association-rule-based approach allows us to investigate point-wise directional relationships between scalars of different variables.

5.3.2 Probabilistic Association Graph

Given a multivariate data set, the scalars and the associated multi-scalar interactions can be represented by a directed graph, where every node $i \in V$ is a scalar x_i and every directed edge $e(i, j) \in E$ is an association rule $x_i \rightarrow x_j$ with the confidence $p(x_j|x_i)$. Essentially, the graph represents information propagation between scalars of different variables, as edges reflect the information flows. We call this the probabilistic association graph $PAGraph(V, E)$. For a system of M variables, since every edge connects two nodes from different variables and no edge connects two nodes from the same variable, a PAGraph is basically a *multipartite graph* whose nodes can be divided into M disjoint sets. Figure 5.2 illustrates examples of PAGraphs with $M = 2, 3, 4$.

Since a continuous scalar field contains an infinite number of values, each value domain is discretized into N bins to represent the entire field. We take the value at the center of



Bipartite-PAGraph Tripartite-PAGraph Quadripartite-PAGraph

Figure 5.2: An illustration of PAGraphs for multivariate data sets (distinct colors are used for nodes from different variables). Left: bipartite-PAGraph for two variables. Middle: tripartite-PAGraph for three variables. Right: quadripartite-PAGraph for four variables.

a bin to represent the scalars that lie within the range of the bin. As a result, the number of nodes $|V|$ is upper-bounded by MN, and the number of edges $|E|$ is upper-bounded by $M(M-1)N^2$. Since many scalar values of different variables do not co-occur in the multidimensional space, $|E|$ is often much smaller than $M(M-1)N^2$. To compute the conditional probability $p(x_j|x_i)$ for a pair of scalars of two different variables, we use 1D histogram to approximate the probability distribution $p(x_i)$ for every variable, and 2D joint histogram to approximate the joint probability distribution $p(x_i,x_j)$ for every pair of variables, and hence we have $p(x_j|x_i) = p(x_i,x_j)/p(x_i)$.

5.3.3 Informativeness and Uniqueness of Scalars

With the PAGraph, it is now possible to identify informative scalars over the multivariate domain, which can infer the existence of other scalars of different variables. From the standpoint of information flows, this is similar to identifying influential individuals in a social network based on their information propagating activities. In this section, we first discuss the analogy between social networks and PAGraphs in terms of two important

factors: influence/informativeness and passivity/uniqueness. Then we describe how to determine the informativeness and uniqueness of scalars in the multivariate domain using our PAGraph.

From Social Networks to Association Graphs

In the field of social network analysis, *influence*, which makes people adapt their behavior, attitudes or beliefs, has been an important factor that directs the dynamics of social media interactions [87]. High popularity does not necessarily infer high influence and vice-versa [120]. The further one's messages are propagated in the network by their connected people, the more influence they may have on others. Equally important is *passivity*, as many people are passive information consumers and do not forward the information to the network [120]. Passivity reflects the barrier to the propagation of messages that is often difficult to overcome. In a brief way, the influence and passivity measures of individuals in social networks can be interpreted as follows:

- **Influence**: an individual of high influence can attract his audience and lead the social interactions.
- **Passivity**: an individual of high passivity is often inert in response to the social interactions.

We observe two factors similar to influence and passivity in the multivariate domain. First, inter-dependencies exist between scalars of different variables, and the scalars that often occur with others actively engage in multi-scalar interactions, and thus have more information in terms of their predictability of the dependent scalars. It is also true that highly frequent scalars are not necessarily the most informative, as they can correspond

to background noise or uninteresting regions. Second, certain scalars may not be easily inferred from other scalars of different variables, and thus are considered unique in the multi-scalar associations. In analogy to influence and passivity in social networks, we define the *informativeness* and *uniqueness* of scalars in the multivariate domain:

- **Informativeness**: a scalar of high informativeness can reliably infer the existence of other scalars of different variables.
- **Uniqueness**: a scalar of high uniqueness can hardly be inferred by other scalars of different variables.

To determine informativeness and uniqueness in multivariate data sets, we propose a multivariate association analysis method named the *Multi-Scalar Informativeness-Uniqueness (MSIU)* model, based on the *Influence-Passivity (IP)* model [120]. Next we introduce the basics of the IP model and its applications in social network analysis, and then describe our MSIU model in the multivariate scenario.

Basics of the Influence-Passivity Model

The IP model was first proposed by Romero et al. [120] to determine the influence and passivity of people in a social network based on the structural properties of the network as well as the diffusion behaviors among people. It utilizes the pairwise association information between people to calculate the relative influence and passivity each person has on the whole network. The IP model lies on the basis of the following relations in a social network:

- A person's influence relies on the number of people he influences as well as their passivity.

- A person's influence relies on how much he attracts the attention of his audience compared to other people.
- A person's passivity relies on the influence of those who he is exposed to but not influenced by.
- A person's passivity relies on how much he rejects other people's influence compared to everyone else.

Given a social network $G(V,E)$ with individuals V and edges E, the weight w_{ij} on an edge $e(i,j) \in E$ represents the ratio of influence that person i does exert on person j to the total influence that i attempts to exert on j. Romero et al. [120] used the ratio of retweets as w_{ij}, which is the number of i's tweets retweeted by j divided by the total number of i's tweets. For every edge $e(i,j) \in E$, the *acceptance rate* is defined as:

$$a_{ij} = \frac{w_{ij}}{\sum_{k:(k,j)\in E} w_{kj}}. \tag{5.1}$$

This metric reflects the dedication j has to i in terms of information propagation. It measures the amount of influence that j accepts from i scaled by the total influence accepted by j from all people in the network. For example, Figure 5.3 shows that j accepts influence from four neighbors including i, and the acceptance rate a_{ij} reflects how much i attracts the attention of j competing with other j's neighbors who may also influence j. On the other hand, the *rejection rate* is defined as:

$$r_{ij} = \frac{1 - w_{ij}}{\sum_{k:(i,k)\in E} (1 - w_{ik})}. \tag{5.2}$$

This metric measures the amount of influence that j rejects from i scaled by the total influence from i rejected by all people in the network. This can also be illustrated using the example in Figure 5.3: i attempts to influence four neighbors including j, and the rejection

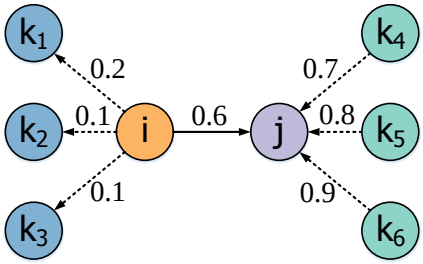


Figure 5.3: An illustration of information propagation in a social network or a PAGraph. The direction of an edge shows the direction of information propagation.

rate r_{ij} assesses how much j rejects i's influence compared to other i's neighbors who may also reject i's influence.

The Multi-Scalar Informativeness-Uniqueness Model

Inspired by the IP model in [120] for social network analysis, we propose the MSIU model to evaluate the informativeness and uniqueness of scalars in the multivariate domain. Table 5.1 shows the correspondence between the IP model in [120] and our MSIU model. In our MSIU model, we take a *PAGraph* as the input graph, which corresponds to a social network in the IP model. Our MSIU model considers *scalars* as nodes, and edges are constructed based on *association rules*, which is analogous to the concept of Twitter users being nodes and retweeting relationships being edges in the IP model. The edge weight in our MSIU model is the *confidence* of the association rule, which corresponds to the ratio of retweets in the IP model. While popularity in the IP model is reflected by the number of followers, it is the probability of occurrence for a scalar in our MSIU model. The relations between informativeness and uniqueness can also be interpreted in the multivariate domain:

Table 5.1: The correspondence between the IP and MSIU models.

	The IP model in [120]	**Our MSIU model**
Input graph	Social network	PAGraph
Nodes	Twitter users	Scalars
Edges	Retweets	Association rules
Edge weight	Ratio of retweets	Rule's confidence
Popularity	Number of followers	Probability
Output	Influence	Informativeness
	Passivity	Uniqueness

- A scalar's informativeness depends on the number of scalars it infers as well as their uniqueness.
- A scalar's informativeness depends on how reliably it infers the existence of other scalars compared to everyone else.
- A scalar's uniqueness depends on the informativeness of those who it is exposed to but not informed by.
- A scalar's uniqueness depends on how much it rejects other scalars' information compared to everyone else.

Given a pair of scalars in two different variables, the amount of information that flows from one scalar x_i to another x_j is x_i's confidence of inferring the existence of x_j, which is measured by $p(x_j|x_i)$. Considering all possible associated scalars, the acceptance rate of x_j

with respect to x_i is defined as:

$$a_{ij} = \frac{p(x_j|x_i)}{\sum_{k:(k,j)\in E} p(x_j|x_k)}. \tag{5.3}$$

The acceptance rate in our MSIU model measures the amount of information that x_j accepts from x_i scaled by the total information accepted by x_j from all its associated scalars in the PAGraph. Take Figure 5.3 as an example: x_j is informed by four other scalars, with confidence $p(x_j|x_i) = 0.6$, $p(x_j|x_{k_4}) = 0.7$, $p(x_j|x_{k_5}) = 0.8$ and $p(x_j|x_{k_6}) = 0.9$, and hence we have $a_{ij} = 0.6/(0.6+0.7+0.8+0.9) = 0.2$. The acceptance rate a_{ij} reflects how much x_i can infer the existence of x_j competing with other x_j's neighbors who may also inform x_j. In this example, although the information that flows from x_i to x_j is 60%, the relative information that x_j accepts from x_i is very small (20%) when considering all the associated scalars of x_j. Such a relative measure is desired in our multivariate scenario as it takes all possible associations into account for quantifying the relative information that flows from one scalar to another.

Given a pair of scalars in two different variables, the amount of information from x_i that is rejected by x_j is the chance of x_j's absence when x_i is observed, which is measured by $p(\bar{x_j}|x_i) = 1 - p(x_j|x_i)$. Considering all possible associated scalars, the rejection rate of x_j with respect to x_i is defined as:

$$r_{ij} = \frac{1 - p(x_j|x_i)}{\sum_{k:(i,k)\in E} (1 - p(x_k|x_i))} = \frac{p(\bar{x_j}|x_i)}{\sum_{k:(i,k)\in E} (p(\bar{x_k}|x_i)}. \tag{5.4}$$

The rejection rate in our MSIU model measures the amount of information that x_j rejects from x_i scaled by the total information from x_i rejected by all its associated scalars in the PAGraph. In the example of Figure 5.3, we see that x_i attempts to inform four other scalars, with rejected information $p(\bar{x_j}|x_i) = 1-0.6 = 0.4$, $p(\bar{x_{k_1}}|x_i) = 1-0.2 = 0.8$, $p(\bar{x_{k_2}}|x_i) = 1-0.1 = 0.9$ and $p(\bar{x_{k_3}}|x_i) = 1-0.1 = 0.9$, and hence we have $r_{ij} = 0.4/(0.4+0.8+0.9+$

$0.9) = 0.13$. The rejection rate r_{ij} assesses how much x_j rejects x_i's information compared to other x_i's neighbors who may also reject x_i's information. In this example, the amount of information that is rejected by x_j from x_i is relatively small (13%) after considering all possible associated scalars of x_i.

Both acceptance rates and rejection rates in our MSIU model are used to obtain the relative informativeness and uniqueness that one scalar has in the entire PAGraph, based on the information flows modeled by the association rules. We employ the iterative scheme in the IP model [120] to compute the informativeness and uniqueness in our MSIU model:

$$I_i \leftarrow \sum_{j:(i,j)\in E} a_{ij} U_j \tag{5.5}$$

$$U_i \leftarrow \sum_{j:(j,i)\in E} r_{ji} I_j \tag{5.6}$$

In the initial state, the informativeness and uniqueness values are set to 1 for every scalar, and are iteratively updated as the information flows between pairs of scalars until convergence. At the end of each iteration, I_i and P_i are normalized to be within $[0,1]$ respectively. Romero et al. [120] have shown that this iterative process converges in tens of iterations for even large graphs that have one million edges.

Essentially, the informativeness and uniqueness are quantified based on how information is propagated between scalars of different variables. From equation (5.5), we see that a scalar x_i is more informative if the following conditions hold: (1) x_i is associated with many other scalars of different variables; (2) the information of x_i is highly accepted by the associated scalars (a_{ij} is high); and/or (3) the associated scalars are unique (U_j is high), which means that x_i is able to infer the existence of scalars that are generally hard to be predicted. Similarly, a scalar x_i is more unique if (1) x_i is associated with many other scalars of different variables; (2) x_i mostly rejects the information of the associated scalars (r_{ji} is

high); and/or (3) the rejected scalars are informative (I_j is high), which means that even the generally informative scalars are not able to infer the existence of x_i.

The complete process of determining the informativeness and uniqueness that each scalar has in the multivariate domain is referred to as the Multi-Scalar-Informativeness-Uniqueness (MSIU) algorithm (Algorithm 2). We now analyze the time complexity of the MSIU algorithm. For a data set of M variables $\times$ K data points, 1D histograms and 2D joint histograms can be computed in $O(MK + M(M-1)K)$ time. Assuming the number of bins in each scalar range is N, the PAGraph can be constructed in $O(M(M-1)N^2)$. Computing the acceptance rates and rejection rates takes $O(|V| + |E|)$ time. The iterative computation of informativeness and uniqueness values takes $O(c(|V| + |E|))$, where c is the total number of iterations. According to Romero et al. [120], c is a small number since the iterative process converges very quickly. Since $|V|$ is upper-bounded by MN, and the number of edges $|E|$ is upper-bounded by $M(M-1)N^2$ ($|E|$ is often much smaller than $M(M-1)N^2$ due to the sparsity of multidimensional data), overall the MSIU algorithm takes $O(MK + M(M-1)K + c(MN + M(M-1)N^2))$ time.

5.4 Association-Guided Exploration Framework

We combine our association analysis method with interactive visualization techniques to develop a framework for exploring scalar-level associations in multivariate data sets. In this section, we describe the design considerations and choices of our system interface, and provide guidelines for visual exploration using our framework. A supplementary video that accompanies this dissertation demonstrates our user interface and the exploration work flow.

Algorithm 2 Multi-Scalar Informativeness-Uniqueness (MSIU)

Input: a multivariate data set V.

Output: informativeness I_i and uniqueness U_i for every scalar $x_i \in V$.

1: Compute probability distribution $p(x_i)$ for each variable, and joint probability distribution $p(x_i, x_j)$ for every pair of variables.
2: Construct a $PAGraph(V,E)$ with E being edges based on association rules of V, and w_{ij} for each $e(i,j)$ being $p(x_i, x_j)$.
3: Compute the acceptance rate a_{ij} and rejection rate r_{ij} for each $x_i \in V$ according to equations (5.3) and (5.4) respectively.
4: Initialize informativeness $I_i = 1$ and uniqueness $U_i = 1$ for each $x_i \in V$.
5: **repeat**
6: **for** each $x_i \in V$ **do**
7: $I_i \leftarrow \sum_{j:(i,j)\in E} a_{ij} U_j$
8: **end for**
9: **for** each $x_i \in V$ **do**
10: $U_i \leftarrow \sum_{j:(j,i)\in E} r_{ji} I_j$
11: **end for**
12: **for** each $x_i \in V$ **do**
13: $U_i \leftarrow \frac{U_i}{\sum_{k\in V} U_k}$
14: $I_i \leftarrow \frac{I_i}{\sum_{k\in V} I_k}$
15: **end for**
16: **until** I_i and U_i converge

5.4.1 User Interface

The user interface consists of three components: the PAGraph view, the multi-faceted PCP view, and the spatial view.

PAGraph View

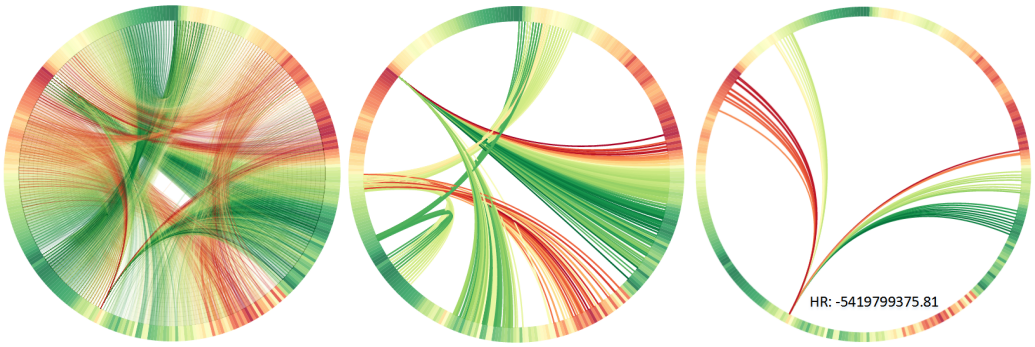
The PAGraph view is to visualize the PAGraphs of confident associations between scalars of different variables. Because a PAGraph typically has over hundreds of nodes and thousands of edges, conventional node-link diagrams can suffer from visual cluttering due to node overlap and edge crossing. An alternative visualization is the adjacency matrix, which shows how nodes are connected together through the intersection of rows and

columns. However, considering the sparsity of multidimensional data as well as the multi-partite property of PAGraph, the PAGraph is often sparse and many display areas will be wasted in the corresponding adjacency matrix visualization. In searching the design space for visualizing directed graphs, we choose a radial graph visualization (Figure 5.4): the scalars are arranged clockwise around a circle in an ascending order of the values, while the entire circle is divided into multiple segments for multiple variables (the scalars that do not have associations are omitted). Associations between scalars of different variables are drawn as curves connecting the associated scalars together. Each scalar node on the circle is colored by the its informativeness in a green-yellow-red order (see Figure 5.1 for the color legend); and the color of an arc is consistent with one of the associated scalars to facilitate visual tracking. The boundaries between the scalars of different variables along the circle are marked by small line segments (see Figure 5.6.1).

In an initial PAGraph view (Figure 5.4a), the top confident associations are visible (10% by default), and users can adjust the edge density of the PAGraph to be shown in a control panel. Users can hover on a scalar node in the outermost circle of the PAGraph to view its confident associations, and a tooltip will pop up for showing the corresponding scalar value in focus (Figure 5.4c).

Multi-Faceted PCP View

With our analysis method, a scalar value (val.) now has three additional facets: informativeness (inf.), uniqueness (uni.) and probability (pro.). A multi-faceted view of scalars can reveal the relationships between the facets and prompt selecting scalars of interest [73]. To this end, we employ parallel coordinates plot (PCP) [61], with the following special design considerations:



(a) (b) (c)

Figure 5.4: Visualization of PAGraphs with different number of associations. (a) Top 10% confident associations; (b) Associations of a few scalars; (c) Associations of one particular scalar.

Rank-based Axes. The density of polylines on a PCP axis is typically not uniform, due to the non-uniform distribution of data values. As a result of such overplotting, users may not be able to select a small portion of scalars of interest in dense areas through brushing [10]. We transform the values on each axis into a *rank space*, where the position of a polyline on an axis shows the *relative rank* of the facet for the corresponding scalar. Since ranks are uniformly distributed on an axis, users can brush a small region for selecting a few scalars of interest (Figure 5.6.2). By default the polylines are colored by their informativeness in a green-yellow-red order (see Figure 5.1 for the color legend).

Juxtaposed Plots. To show the multi-ranks of scalars of multiple variables simultaneously, we employ the juxtaposition design [43], using multiple side-by-side PCPs — each PCP corresponds to multi-faceted scalars of one variable (Figure 5.6.1). Such juxtaposed plots allow users to focus on selecting an interesting value range in one variable, while still maintaining the information of the multi-ranks in other variables for switching the selection.

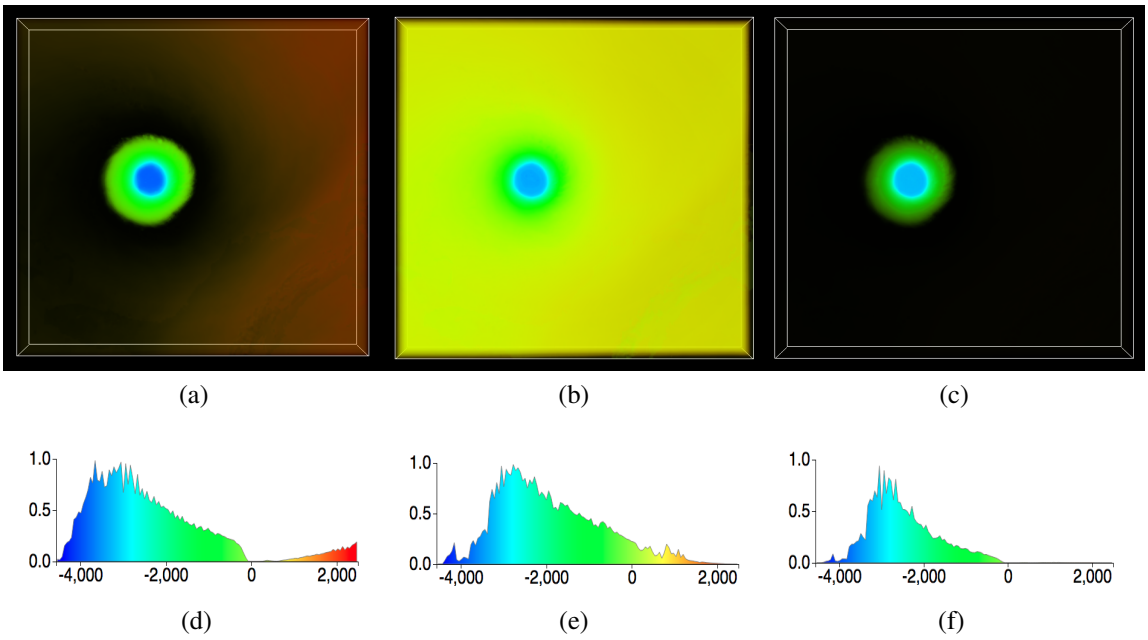


Figure 5.5: Volume rendering of the PRE variable in the Hurricane Isabel data set based on average associations with the scalars of the HR and OH variables. Informative PRE volume (a) with its transfer function (d); Unique PRE volume (b) with its transfer function (e); Informative and unique volume (c) with its transfer function (f).

Cross-linked Views. In addition to standard brushing and linking across axes [10], our design also supports brushing and linking across different variables — when users select scalars in the PCP of one variable through brushing, the top associated scalars of other variables are extracted from the PAGraph (10% by default), and are simultaneously highlighted in the other PCPs. Furthermore, these associations are also highlighted in the PAGraph view (Figure 5.6.2). The values of the selected scalars are shown in a separate window so that users can understand and refine their selection.

Spatial View

While the PAGraph view and the PCP view together serves as the interactive interface for users to select scalars of interest and view their associations in detail, the spatial view provides users with the most direct and intuitive view of the data through volume rendering (Figure 5.6.3b and 5.6.3c) and isosurface visualization (Figure 5.6.3a). To design an automatic insightful transfer function that saves users' effort from the tedious and time-consuming manual specification, while also revealing the associations between scalars of different variables, we propose the following two schemes:

Transfer Function with Average Associations. The informativeness and uniqueness of a scalar together reveals the association information about scalars of other variables in an average sense. Therefore, the transfer function that reflects average scalar-level associations takes one of the following options: $opacity(informativeness(x))$, $opacity(uniqueness(x))$, or $opacity(informativeness(x) \times uniqueness(x))$. The corresponding volumes determined by these transfer functions are referred to as *informative volume*, *unique volume*, and *informative and unique volume* respectively, as illustrated in Figure 5.5.

Transfer Function with Specific Associations. When users select a scalar of interest in the interface, in addition to showing the corresponding isosurface visualization (Figure 5.6.3a), the *associated volumes* highlight the scalars of other variables that are associated with the selected scalar. Thus, for a given scalar x, the transfer function of its associated volume Y is set to $opacity(p(y|x))$. Figure 5.6.3b and 5.6.3c present an example of two associated volumes with respect to a selected scalar.

Empirically, we use $opacity(f(x)) = [f(x)]^2$ to emphasize the relatively high informative/unique scalars or the relatively confident associations. The color transfer function $color(x)$ maps the scalar values in a blue-green-red order, as shown in Figure 5.1.

5.4.2 Exploration Guidelines

For a given multivariate data set, we assume that users already have some pre-selected variables in mind, as is often the case for domain scientists in real-world applications [14]. The task to address in this work is to explore scalars of interest in the given variables based on scalar-level associations. We follow the visualization mantra "Overview first, zoom and filter, then details-on-demand" by Ben Shneiderman [124] to guide the exploration process:

Overview first. The initial PCP and PAGraph views provide an overview of the multiple facets of scalars and their associations (Figure 5.7.1). Users can select a variable from the PCP labels to initiate the spatial view with the informative/unique volume of the variable (Figure 5.7.2).

Zoom and filter. To understand how scalars of other variables are associated with a particular scalar, users can select a scalar by brushing the PCP view and see the corresponding associations in the PAGraph view. For instance, users can first brush the Inf./Uni. axis to selected a few informative/unique scalars, and then brush the Pro. axis to narrow down the focus to a particular scalar of interest (illustrated in Figure 5.7.3 as well as in the supplementary video).

Details-on-demand. To gain insights into the spatial relationship of the selected scalar and its associated scalars of other variables, the spatial view shows the isosurface visualization of the selected scalar and its associated volumes of the other variables (Figure 5.7.4).

5.5 Results

We demonstrate the effectiveness of our methods through experiments on three representative multivariate data sets: Hurricane Isabel, Ionization Front Instability, and Turbulent Combustion. The experiments were conducted on a desktop machine with an Intel

core i7-2600 CPU, 16 GB of RAM and an NVIDIA GeForce GTX 560 GPU with 2GB texture memory. The interactive exploration process in each experiment is demonstrated in the supplementary video that accompanies this dissertation.

5.5.1 Case Study 1: Hurricane Isabel Data Set

The Hurricane Isabel data set is an atmospheric simulation created by the Weather Research and Forecast model, courtesy of the National Center for Atmospheric Research and the U.S. National Science Foundation. The resolution of the grid is $250 \times 250 \times 50$ at each time step. To investigate the relationships between pressure, humidity and wind speed in the hurricane simulation, we selected the Pressure (PRE), Water Vapor Mixing Ratio (QVA) and Wind Velocity Magnitude (VEL) variables at time step 18 for our experiment on this data set.

Figure 5.6.1 presents the initial PCP view and PAGraph view of the selected variables. We started the exploration by selecting the PRE variable and examine the resulting informative/unique volumes, as shown in Figure 5.5. We can see that the hurricane eye and eyewall structures are highlighted by the informative and unique PRE volume in Figure 5.5c, which has low air pressure as illustrated in Figure 5.5f.

To understand how the PRE scalars interact with the scalars of the other variables QVA and VEL, we brushed the initial PCP view in Figure 5.6.1 to select one informative scalar PRE = -2674.17 in Figure 5.6.2. The linked PAGraph view in Figure 5.6.2 highlights the confident associations starting from a single node (PRE = -2674.17). From the linked PCP of VEL in Figure 5.6.2, we observe that a majority of the associated VEL scalars have high values. Meanwhile, we found that the selected PRE scalar is associated with high/medium QVA values that represent a decent amount of water vapor. When it comes to the spatial

view in Figure 5.6.3, we found that the isosurface PRE = -2674.17 in Figure 5.6.3a corresponds to the hurricane eyewall. The associated QVA volume in Figure 5.6.3b presents the long bands of rain clouds that spiral inward to the hurricane eye, and the associated VEL volume in Figure 5.6.3c highlights strong winds near the hurricane eyewall. In this sense, this informative PRE scalar with its associated QVA and VEL scalars presents a joint feature from multiple variables — the hurricane eyewall with strong wind and spiral rainbands.

As a follow-up comparison study, we select an uninformative scalar PRE = 141.61 in Figure 5.6.4. We see that this scalar mostly interacts with low VEL values and low QVA values. The green PCP polylines in Figure 5.6.4 imply that these associated scalar values are also uninformative. The spatial view in Figure 5.6.5 further explains such associations: the isosurface PRE = 141.61 spreads out the entire spatial domain except around the hurricane eye (Figure 5.6.5a), which is associated with mostly low water vapor on the continent (the dark blue regions in Figure 5.6.5b) and weak wind faraway from the hurricane eye (the dark blue regions in Figure 5.6.5c).

5.5.2 Case Study 2: Ionization Front Instability Data Set

The Ionization Front Instability data set is a simulation created by Mike Norman and Daniel Whalen for investigating the effects of instabilities where radiation ionization fronts scatter around primordial gas in the formation of galaxies. Ionization front is a transition region where interstellar gas changes from a mostly neutral state to a mostly ionized state, which involves the abundances of many chemical species such as hydrogen and helium.

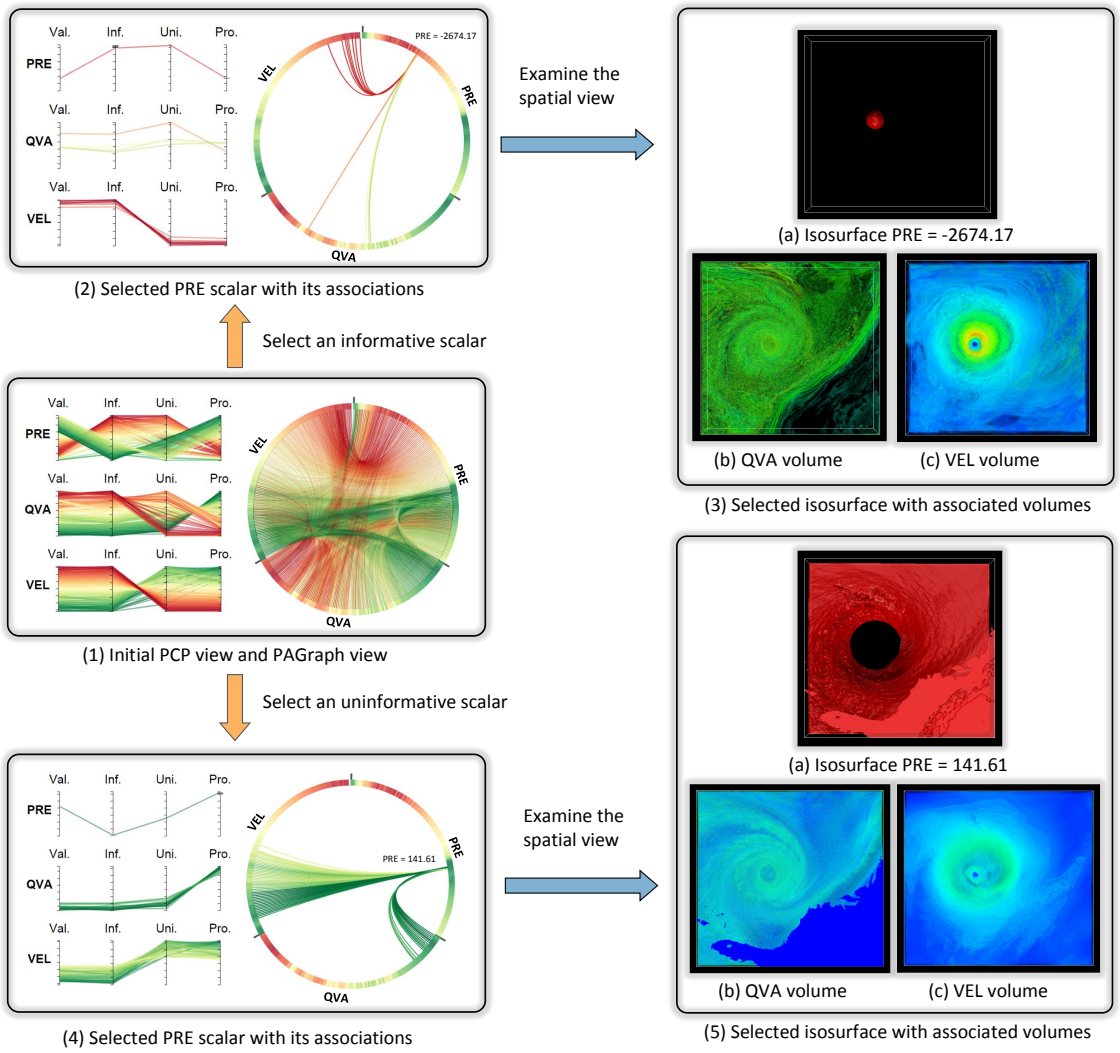


Figure 5.6: Experiments on the Hurricane Isabel data set. See Section 5.5.1 for details about the exploration process.

The resolution of the grid is $300 \times 124 \times 124$ at each time step. To investigate the dependency of the ionization process with respect to temperature, we selected the variables of H, H^+ (HP), He (HE), and Temperature (TEM) and time step 100 in our experiment.

For this data set, the initial PCP view and PAGraph view are presented in Figure 5.7.1. We initiated our exploration with the informative and unique TEM volume in Figure 5.7.2, determined by the transfer function $opacity(informativeness(x) \times uniqueness(x))$. We found that the TEM scalars around 10000K are highly informative and unique. We then select an informative and unique scalar TEM = 10588.16 in the PCP view in Figure 5.7.1. From the updated PCP view and PAGraph view in Figure 5.7.3, we can see that many scalars are associated with this temperature value. To examine how the selected chemical species react at this temperature 10588.16K, we resort to the spatial view of the associated volumes in Figure 5.7.4b, 5.7.4c and 5.7.4d. It is obvious that the regions overlapping the isosurface TEM = 10588.16 in Figure 5.7.4a are mostly highlighted, particularly in the associated H and HP volumes. According to the studies of hydrogen ionization [131], 10000K is roughly the temperature where a majority of neutral hydrogen become ionized hydrogen. Furthermore, we also explored uninformative TEM scalars from the PCP view in Figure 5.7.1, which have very high temperature values. No confident associations (top 10%) were found between those uninformative scalars and the H/HP/HE scalars (the resulting PCP and PAGraph views are not shown due to the lack of space).

5.5.3 Case Study 3: Turbulent Combustion Data Set

The Turbulent Combustion data set is a combustion simulation produced by Dr. Jacqueline Chen at Sandia Laboratories through US Department of Energy's SciDAC Institute for Ultrascale Visualization. The resolution of the grid for each time step is $240 \times 360 \times 60$. The Mixture Fraction (MIX) variable in this data set, ranging from 0 (pure oxidizer) to 1 (pure fuel), signifies the proportion of fuel and oxidizer and generally reflects the characteristics of the flame: when the chemical reaction rate is greater than the turbulent mixing rate,

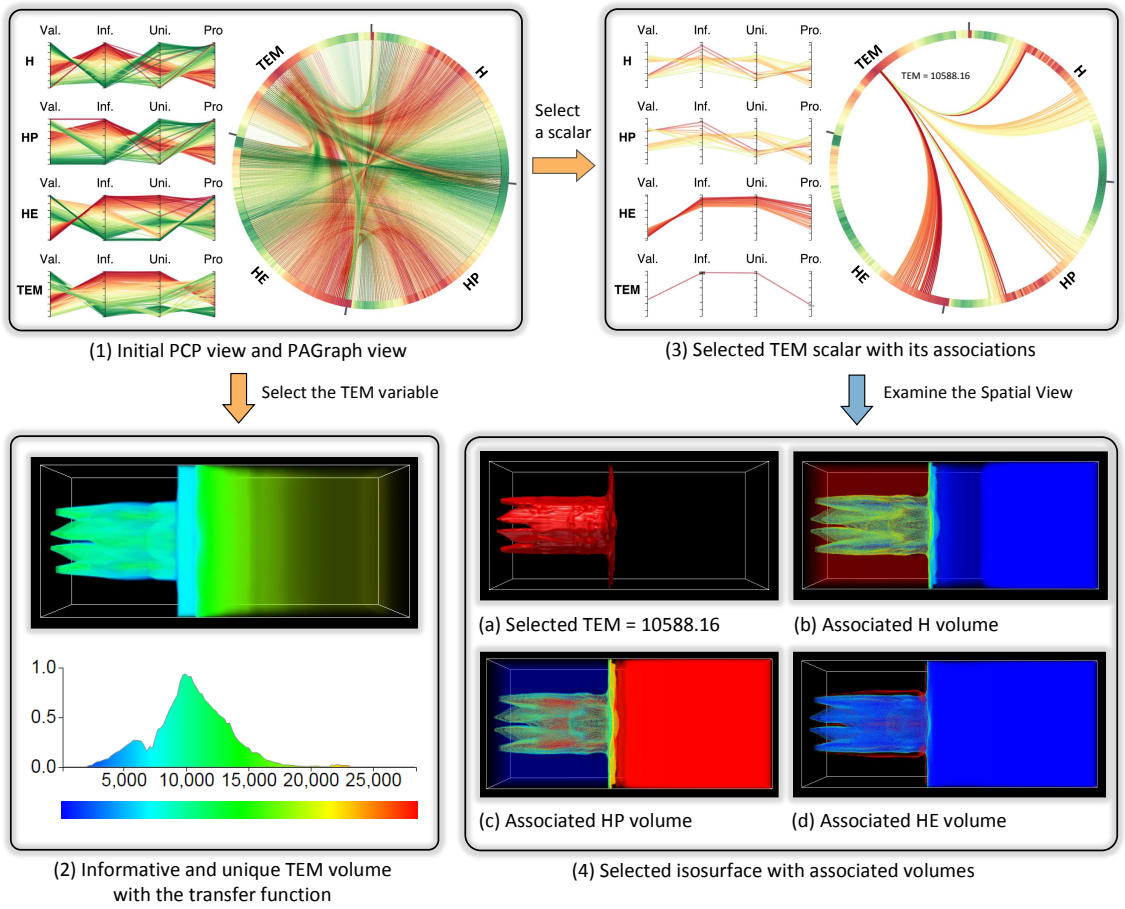


Figure 5.7: Experiments on the Ionization Front Instability data set. See Section 5.5.2 for details about the exploration process.

a complete burning is indicated (the scalar MIX = 0.42 corresponds to a complete burning as reported in [3]); weak or extinguished burning occurs when the turbulent mixing rate exceeds the chemical reaction rate. As the characteristics of local combustion also depends on the Heat Release Rate (HR) and Mass Fraction of the Hydroxyl Radical (OH) from the neighboring flame elements, we are interested in exploring scalar-level associations of the MIX, HR and OH variables. We have selected time step 65 to conduct our experiment.

The initial PCP view and PAGraph view are presented in Figure 5.8.1. From the PCP view, we observe that the most informative MIX scalars lie in the middle range of MIX. Since flames typically burn in a balanced mixture of fuel and oxidizer (either pure oxidizer or pure fuel will result in extinction of reaction), we expect these informative MIX scalars correspond to the burning flame regions, which strongly interact with scalars of the HR and OH variables. The informative MIX volume in Figure 5.8.2 illustrates the spatial locations of such flame structures.

To investigate the associations between the MIX scalars and the scalars of the other variables HR and OH, two specific MIX scalars that have contrast informativeness have been selected in Figure 5.8.3 and 5.8.4 respectively. The linked PAGraph view in Figure 5.8.3 highlights a lot of associations with MIX = 0.4998, which we expect to represent a burning flame. From the linked PCP view of OH in Figure 5.8.3, we see that a majority of the associated OH scalars have high values. This is consistent with the fact that the highly convoluted flame regions often have rapid radical diffusion rates [3]. Similarly, we found that the scalar MIX = 0.4998 is associated with many low (negative) HR values that represent high heat release rates. In contrast, an uninformative scalar MIX = 0.0062 in Figure 5.8.4, which we expect to be a weakly burning region due to the lack of fuel, has only a few associations. From the linked PCP views of OH and HR in Figure 5.8.4, we found that the associated scalars are low OH values and high HR values. This indicates that weakly or extinguished burning regions often have low radical diffusion rates and low heat conduction rates.

The spatial views in Figure 5.8.5 and 5.8.6 confirmed our findings: the associated HR/OH volume mostly overlaps with the flame structure represented by the isosurface MIX = 0.4998, while the isosurface MIX = 0.0062 presents two outer layers in the spatial

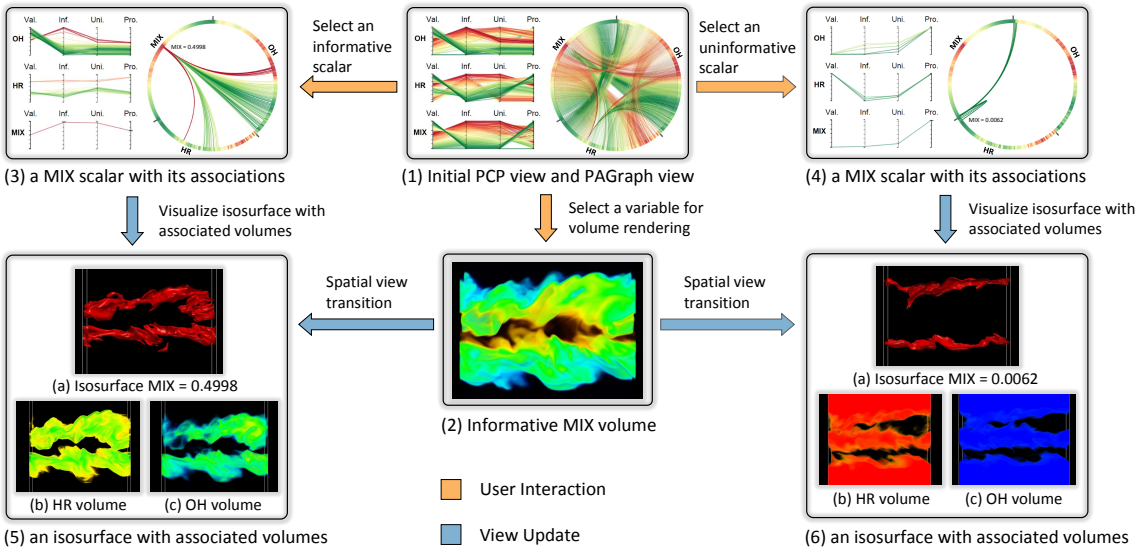


Figure 5.8: An illustration of our association-guided exploration framework using the Turbulent Combustion data set. Starting from the initial PCP/PAGraph view (1), users can select one variable to see the informative volume rendering in the spatial view (2), or select specific scalars to see their associations in the PCP/PAGraph view (3, 4), together with the isosurface visualization and associated volumes in the spatial view (5, 6).

domain that is far away from the strong flame region, and the associated HR/OH volumes highlight the regions that are either outer layers (where oxidizer is too high) or inner layers (where fuel is too high).

5.6 Discussion

The MSIU model brings a novel perspective to visual exploration of multivariate data sets. Previous correlations methods include correlation coefficients, mutual information, gradient-based measures, etc. Compared to these approaches, our method has two major differences: (1) the MSIU model takes into account the associations of scalars, which are the more specific relationships between variables. These scalar-level associations reveal

complex interactions between scalars of different variables, and can be used to classify the scalars in terms of their informativeness and uniqueness; (2) our approach can easily cope with more than two variables, as the MSIU model constructs a multi-partite association graph that captures the relationships of scalars in the given variables. In contrast, existing correlation approaches mostly focus on the relationships of two variables, and the extension to more than two variables is often non-trivial.

The association graph with the informativeness and uniqueness learned from the MSIU model is used for volume rendering that encodes the relationships between scalars in one variable and scalars of the other variables. This unique property differentiates our method from many existing multi-dimensional transfer function design approaches, as they usually map a multidimensional vector to a single color and opacity for describing all the variables, sometimes with additional derived variables. Consequently, it can be difficult to understand the different contributions of individual variables regarding the multivariate interactions. Our association-aware transfer functions, encoding the average or specific associations, are automatically created for each variable respectively. Therefore, the scalars in the spatial view have their intrinsic physical properties based on the semantics of the underlying variable.

Our method works well for multivariate data sets where scalars of different variables co-occur, as shown in the three case studies. As stated in Section 5.4.2, we assume that users already have some pre-selected variables in mind prior to the exploration in real world scenarios [14], and our method has been applied to several variables of interest in each case study respectively. Although one can consider all the variables simultaneously, it can be very difficult to interpret the relationships between a scalar and scalars of a large number of variables. In contrast, pair-wise association or a small number of multivariate

Table 5.2: Average runtime (in seconds) for the MSIU algorithm.

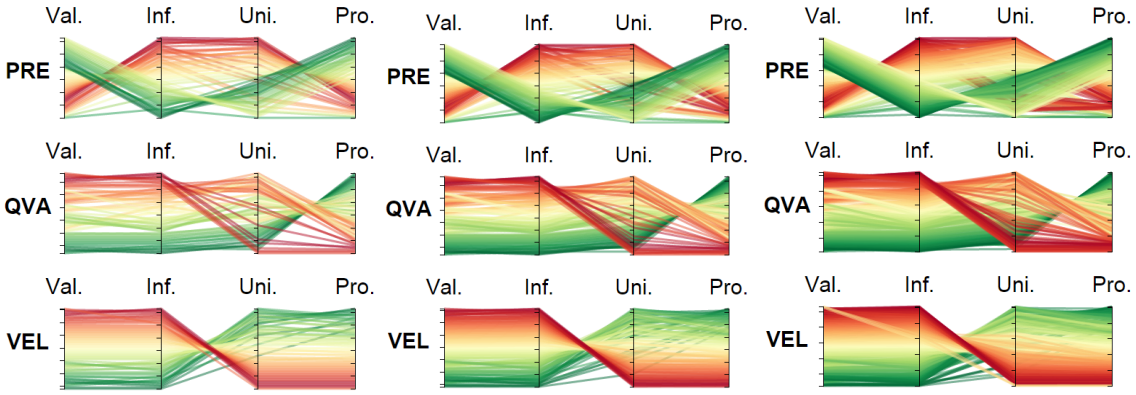
	64 bins	128 bins	256 bins
Hurricane Isabel	1.72	2.62	11.95
Ionization Front	4.33	6.99	14.73
Combustion	2.43	3.40	10.61

associations are more intuitive and easier to understand. The trade-off between multivariate complexity and interpretability needs to be better understood in the future.

All the computations in the MSIU algorithm are done in the preprocessing stage prior to the exploration process. One parameter in the MSIU algorithm is the choice of bin size N, which is related to the discretization of scalar values and the calculation of probability distributions. Table 5.2 reports the computational performance of the MSIU algorithm with different number of bins. We also examined the relative ranks of informativeness and uniqueness when using different numbers of bins. It is observed that a majority of the relative ranks of informativeness and uniqueness remained the same for different numbers of bins with minor variations in all three data sets. Due to space constraint, we only report the comparison results for the Hurricane Isabel data set (Figure 5.9). While we used 128 bins in this work, in the future we would like to experiment with non-uniform discretization as well as multi-level discretization.

5.7 Summary

In this chapter, we present a novel association analysis method to guide visual exploration of scalar-level associations in multivariate data sets. We model the directional



(a) 64 bins (b) 128 bins (c) 256 bins

Figure 5.9: The ranks of informativeness, uniqueness and probability with different number of bins for the Hurricane Isabel data set.

interactions between scalars as information flows based on association rules. We introduce the concepts of informativeness and uniqueness to describe how information flows between scalars of different variables and how they are associated with each other in the multivariate domain. Based on scalar-level associations represented by a probabilistic association graph, we transform the problem of identifying informative and unique scalars into a network analysis problem of finding influential and passive people in social networks, and propose the Multi-Scalar Informativeness-Uniqueness (MSIU) algorithm. We develop an exploration framework with multiple interactive views to explore the scalars of interest with confident associations in the spatial domain, and provide guidelines for visual exploration using our framework. We demonstrate the effectiveness of our approach through case studies in three application domains.

Chapter 6: SocialBrands: Visual Analysis of Public Perceptions of Brands on Social Media

6.1 Motivation

Public perceptions of a brand is critical in determining its performance, as these perceptions influence people's brand preferences and purchase intentions. Social media allows people to share opinions and experiences freely, and offers a great opportunity for companies to assess and construct brand perceptions. As a result, many analytics tools have been designed to assess public perceptions through analyzing trending topics [166], sentiments [102] and emotions [173] on social media. However, existing tools lack a comprehensive framework to assess public perceptions, and many analytics results related to sentiments and emotions can rarely be applied to the daily tasks of domain users (i.e., brand managers) for understanding and managing people's perceptions on brands.

When developing and managing brand perceptions, brand managers often use *brand personality* [1, 104], the most well-established framework to measure brand perceptions in marketing literature. The framework includes a comprehensive set of human characteristics associated with brands. For example, IBM is considered to be *old* and *competent*, while Apple is considered to be *young* and *cool*. In marketing practice, brand managers carefully define the *intended brand personality* that they would like customers to perceive,

and invest extensive resources into brand-related marketing activities to reinforce such perceptions. Nevertheless, it is profoundly challenging to develop and maintain the intended brand personality, as there often exist gaps between the intended brand personality and consumers' actual *perceived brand personality*. Thus, one important task for brand managers is to assess the perceived brand personality in their daily practice. So far brand managers mostly rely on surveys to collect descriptive ratings on multiple brand personality scales to assess brand personality. Unfortunately, conducting surveys is time-consuming and labor-intensive, which makes it difficult to assess brand personality frequently. This problem becomes even more challenging when assessing brand perceptions from multiple perspectives, as perceptions of a brand can be influenced by multiple factors such as typical users and employees of the brand as well as its official marketing messages [113]. More importantly, with survey ratings, brand managers are often incapable of understanding and explaining the reason behind the perceived brand personality traits as well as their relations to multiple driving factors.

To address these challenges, in this chapter[1], we present *SocialBrands*, a novel visual analytic tool for brand managers to assess and analyze public perceptions of brands on social media.

The key contributions of SocialBrands are three-fold.

- A computational approach that is designed to assess brand personality from three driving factors of user imagery, employee imagery and official announcement on social media, and construct an evidence data explaining the association between brand personality and its driving factors.

[1]Major portions of this chapter were previously published in Liu et al. [95].

- A visualization design that conveys an integrated sense of multi-dimensional brand personality with visual evidence and related details. It enables visual explanation of how the perceived brand personality can be derived from the driving factors in social media, and visually highlights varying contributions of social media factors on different personality traits.

- A visualization design that enables a contextual understanding of the marketplace of many brands. It facilitates perception-based market segmentation through visual summarization of the distribution of brands over different personality dimensions and the clusters of brands.

We evaluated the usefulness and usability of SocialBrands through interviews and surveys with brand managers in an enterprise context. The evaluation shows that SocialBrands enabled brand managers to assess and analyze brand perceptions on social media, identify gaps between the intended and the perceived brand personality, understand multiple factors that drive brand personality, gain insights from successful brands through visual comparison, and discover perception-based market segments. Based on our studies, we discussed the design lessons for developing future brand analysis systems.

6.2 Background

This section introduces the background knowledge on brand personality and social media visual analytics.

6.2.1 Brand Personality

A company can significantly enhance its performance and consumer loyalty by building a strong and differentiated brand. A brand is essentially an aggregation of the name,

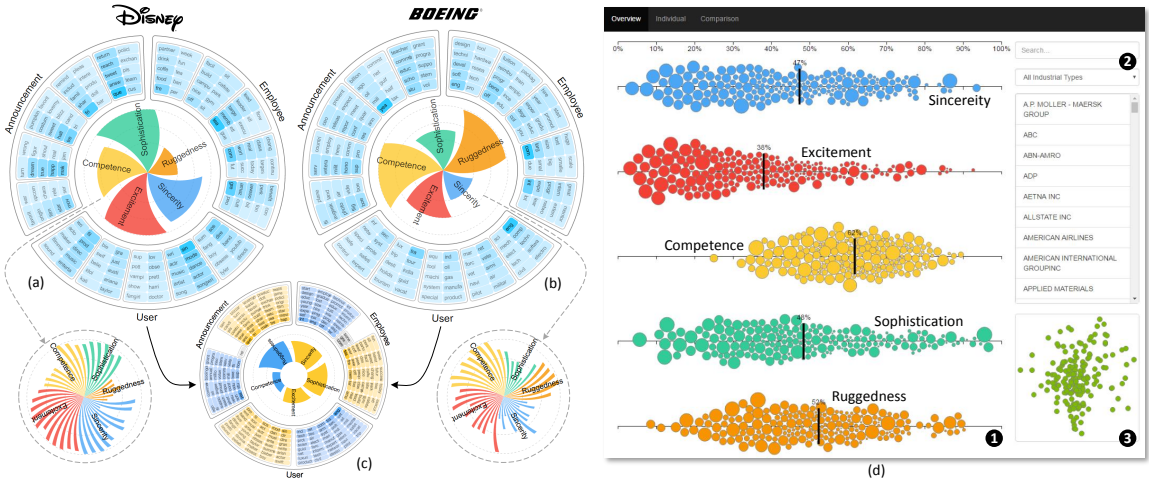


Figure 6.1: SocialBrands illustrates brand perceptions on social media with our designed visualizations. (a,b) The BrandWheels of two brands "Disney" and "Boeing", each illustrates a brand's perceived personality (in 5 broad traits or 42 subtraits) with visual evidence and related details from three social media factors. (c) The Comparative BrandWheel highlights the similarities and differences of two brands in their perceived personalities and topic discussions on social media. (d) The Overview of brand perceptions: (1) BrandSediments visually summarize of the distribution of brands over different personality traits and the clusters of brands; (2) search and filtering widgets; (3) MDS embedding of brand perceptions.

design, symbol and all experiences that distinguishes one company's product from those of others [72]. Brand perceptions are consumers' interpretations of a brand, which can be shaped by functional experiences (e.g., quality and reliability) as well as emotional experiences (e.g., making one feel better or making one's life easier). Brand managers need to comprehensively understand how customers perceive their brand in specific market segments, particularly compared with other competitive companies. In marketing practice, the most well established framework for assessing brand perceptions is ***brand personality***, which is reflected by human-like features that are strongly associated with a brand [104].

Brand personality can be characterized by traits that uniquely identify a brand [72]. *Brand personality scales* [1] serve as a reliable, valid and generalizable measure for assessing brand personality in marketing literature. They consist of 5 broad personality traits: *Sincerity*, *Excitement*, *Competence*, *Sophistication* and *Ruggedness*, which are characterized by 42 subtraits, as shown in Figure 6.2. These personality scales measure how descriptive a trait is of a brand: 0% means not at all descriptive, and 100% means extremely descriptive (e.g., "20% sincerity" means sincerity is slightly descriptive on a $0-100\%$ scale).

Traits	**Subtraits**
Sincerity	down-to-earth, family-oriented, small-town, honest, sincere, real, wholesome, original, cheerful, sentimental, friendly
Excitement	daring, trendy, exciting, spirited, cool, young, imaginative, unique, up-to-date, independent, contemporary
Competence	reliable, hard-working, secure, intelligent, technical, corporate, successful, leader, confident
Sophistication	upper-class, glamorous, good-looking, charming, feminine, smooth
Ruggedness	outdoor, masculine, western, tough, rugged

Figure 6.2: Brand personality scales [1], consisting of 5 broad traits (left column) and 42 subtraits (right column).

Brand personality analytics has received considerable attention in brand management and marketing. Brand managers carefully define the *intended brand personality* that they

would like consumers to perceive, and invest extensive resources into brand-related marketing activities to reinforce such perceptions. However, successful formulation and implementation of brand personality is often a difficult challenge, as there often exist gaps between the intended brand personality and consumers' actual *perceived brand personality*. Thus, one important task for brand managers is to assess the perceived brand personality in their daily practice. Public perceptions of brand personality can be possibly influenced by at least three driving factors: *User Imagery*, *Employee Imagery*, and *Official Announcement* [52]. *User Imagery* are human characteristics associated with a brand's typical users. *Employee Imagery* are human characteristics associated with the employees of a company. *Official Announcement* refers to marketing messages that are designed specifically to engage consumers for brand loyalty and brand awareness.

6.2.2 Social Media Visual Analytics

Social media has transformed the way people market their businesses. Considerable research efforts have been made to support visual analysis of social media data. Dork et al. [28] introduced visual back-channel as a way of following and exploring online conversations about large-scale events. Cuvelier and Aufaure [25] proposed topographic networks of tags, representing a tag cloud with a topographic metaphor to highlight the most important concepts found for a given search on Twitter. Hansen et al. [50] presented a process model to analyze and visualize social media data. Whisper [17] traced and visualized the process of information diffusion in social media using a sunflower metaphor. Liu et al. [89] created a monitoring tool that preserved the user's mental map of streaming tweet clusters. Kraft et al. [81] extracted structured representations of Twitter events and visualized

key event indicators from Twitter stream. Google+ Ripples [155] showed social media interactions on Google+ with a mix of node-and-link and circular Treemap metaphors. Xu et al. [166] analyzed topic computation on social media by integrating ThemeRiver with storyline style visualization. Liu et al. [91] developed a visualization system to analyze Twitter users' influence and passivity conditioned on specific themes. Mahmud et al. [99] supported visual analysis of Twitter users' attitudes towards brands. In contrast with the above tools, we propose a computational approach that derives perceived brand personality from social media data, and present intuitive designs for brand managers to assess perceptions of brands.

With recent advances in human personality and emotion analysis, researchers have developed new visualization tools for exploring analysis results. TwitInfo [102] employed a timeline-based display to highlight peaks of high tweet activity with the associated text sentiment. WeFeelFine [70] collected emotion-related keywords from social media texts for visual search and exploration. PEARL [173] supported multi-dimensional emotion analysis with a timeline-based visualization. FluxFlow [172] combined anomaly detection with thread glyphs and timelines for exploring anomalous information spreading on social media. KnowMe and ShareMe [48] studied the effectiveness of deriving personality traits from social media data and how users shared their personality traits using a multi-view visualization. VeilMe [158] employed a genome representation to visualize personality portraits for privacy configuration. However, existing tools are not designed to understand and assess public perceptions of brands, and many tools related to sentiments and emotions can rarely support the daily tasks of brand managers to analyze and manage people's perceptions on brands. In this work, we applied marketing theories to develop a new visual analytic system for brand personality analytics.

6.3 Domain Problem Characterization

To distill the domain problems and typical tasks for brand management, we worked closely with domain experts such as brand managers from multiple marketing departments in a large international information service corporation. Their daily practice of brand management includes planning, developing, and directing marketing efforts for establishing brand personality, and differentiating it from other competitive brands. To validate their marketing strategies, they want to know how the intended brand personalities are perceived by the targeted audience via various channels including social media. We conducted a series of interviews with the domain experts, and summarized a list of analysis tasks to characterize the domain problem and identify the challenges faced by the target users.

T1. *How to efficiently assess perceived brand personality on social media?* Prior to our work, brand managers mostly used surveys to collect descriptive ratings on brand personality scales to assess brand personality. However, conducting surveys is time-consuming and labor-intensive. It is significantly difficult to assess brand personality frequently. A computational approach that derives perceived brand personality from social media data is urgently needed by the domain experts.

T2. *How to relate a brand's perceived personality with multiple social media factors? Which is the driving factor that contributes most to the perceptions of a certain personality trait?* Brand managers are interested in analyzing the association between social media factors and perceived brand personality. The identification and analysis of varying effects of factors on different personality traits are crucial for them to make corresponding strategic decisions on a specific social media platform.

T3. *What are the reasons that a brand is strong on a certain trait? How to interpret and explain the formation of a certain perception from linguistic footprints on social media?* Brand managers want to determine the reasons behind the perceived brand personality traits. They are concerned with the popular opinions and discussions on social media that have influenced brand perceptions. Linguistic evidence should be provided to facilitate experts' analytical reasoning in understanding and explaining the trait modeling results.

T4. *What are the differences in perceived brand personality of related brands?* Brand managers want to analyze the differences between multiple brands to identify the strengths and weaknesses of a company based on their winning and losing personality traits. This task is critical to evaluating their daily practice as good marketing strategies should differentiate a company from its major competitors in the marketplace. It can also help them identify a set of successful exemplars to improve their existing marketing strategies for achieving better perceptions.

T5. *How to identify market segments within a set of brands from the perspective of perceived brand personality?* Brand managers are encountering with hundreds of brands including their competitors and collaborators in their daily practice. Groups of brands with similar perceived personality are of particular interest for their analysis. Identifying such personality-based market segments can facilitate the domain experts in designing and implementing marketing plans to address specific demands of the target segment.

6.4 A Computational Approach for Brand Perception Analytics

This section presents a computational approach that derives brand personality traits from social media data (G1). We first introduce a multi-level model for computational

brand personality analytics, and then describe how to derive a brand's personality traits from linguistic footprints on social media.

6.4.1 A Multi-Level Model for Brand Personality Analytics

We base our approach on social computing studies [134, 168, 169] to assess perceived brand personality on social media. Computationally, a trait of a brand's personality can be modeled as a linear combination of relevant semantic features:

$$T_i = g_i(F_j(W_k)), \tag{6.1}$$

where g_i is a linear function mapping semantic features to a personality trait, F_j is a function mapping words to semantic features, and W_k is a word that is correlated with trait T_i and appears in the text footprints. To represent the multi-level relationships between personality traits, semantic features, words, and texts, we use a compound graph defined as:

$$TKG = (T, K, g). \tag{6.2}$$

The personality traits of a brand's personality profile are represented as a *personality tree*, $T = (V_T, E_T)$, where a set of nodes V_T denotes the traits, and a set of edges E_T denotes the relationships between the basic traits and their subtraits. The linguistic evidence used to derive a trait, including related semantic features, words, whole texts, and weighted links among them, is represented as an *evidence network*. The evidence network is essentially a k-partite graph, $K = (V_K, E_K), V_K = K_1, K_2, ..., K_k$, where K_l is a set of nodes representing a set of evidence at a particular level l (e.g., word level), and E_K is a set of edges representing the relationships between the nodes. A mapping function, g, links the linguistic evidence to the derived personality traits, $g(n_K) \rightarrow n_T$, where $n_K \in V_K, n_T \in V_T$.

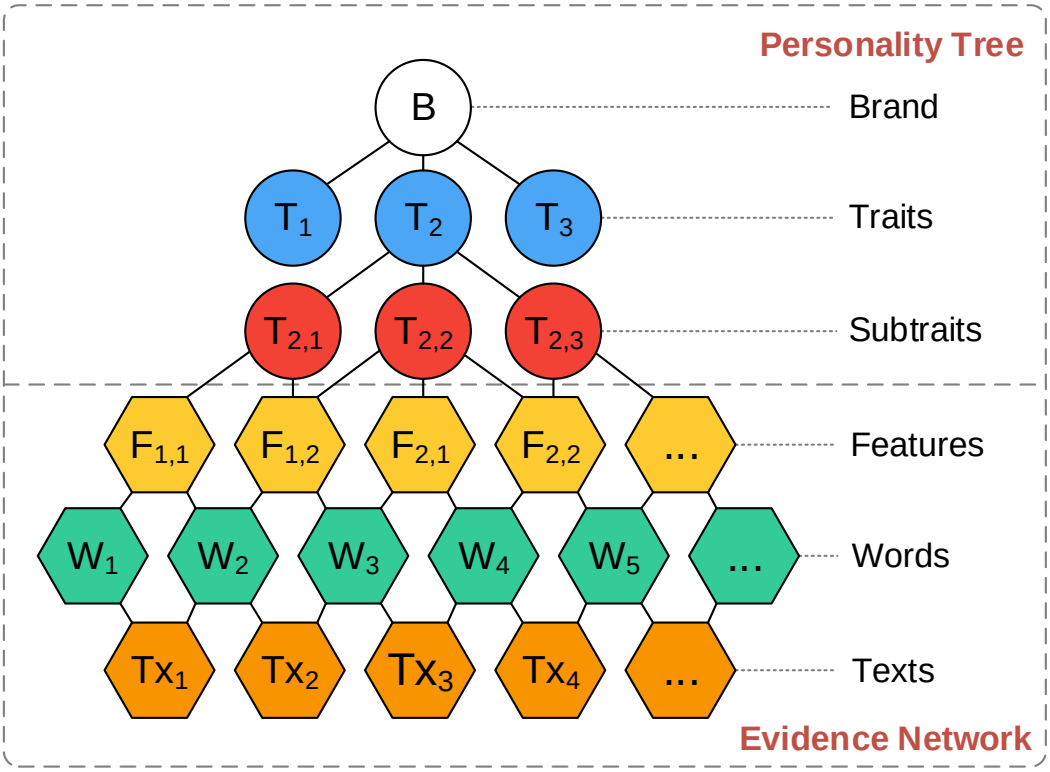


Figure 6.3: The multi-level model of a brand's personality traits (represented as a personality tree) with associated linguistic evidence (represented as an evidence network). This model is used to design a computational workflow to derive brand personality from social media linguistic footprints.

The compound graph view of this multi-level data model is illustrated in Figure 6.3. Given a brand B, its personality tree has several broad traits T_i, which are characterized by subtraits $T_{i,j}$ (the j-th subtrait of the i-th trait). The subtraits are linked with a set of linguistic features $F_{i,j}$ (the j-th feature of the i-th social media factor) in the evidence network. These features are linked with a set of representative words W_i, which are extracted from a set of text documents Tx_i. The key property of this model is that the nodes at one level are derived and linked from the nodes at the level below. This multi-level property

motivates us to design a computational workflow to derive brand personality from social media linguistic footprints by (1) collecting multi-faceted linguistic footprints on social media, (2) extracting semantic features from linguistic texts and words, and (3) mapping semantic features to personality traits. Next we describe the key steps of this computational workflow.

6.4.2 Multi-faceted Social Media Data Collection

In total, we collected data for 219 well-known brands in various industrial sectors such as automobiles, electronics, restaurants, clothing, and financial services. For each brand, we collected data from three facets of linguistic footprints on social media to account for the three known factors that drive brand perceptions [52]: *User Imagery*, *Employee Imagery*, and *Official Announcement*, as follows.

User Imagery A brand's followers on social media are very likely to use and like the particular brand. We considered a set of brand followers as *User Imagery* represented on social media. For each brand, we collected its followers' self-description on Twitter [149], which is a short description in a follower's public profile. Overall, we obtained 1,996,214 brand follower descriptions.

Employee Imagery Glassdoor [41] is an online social media where employees and former employees anonymously review companies and their management. The reviews often contain statements about working conditions, company culture and management styles, which are used to describe *Employee Imagery*. We collected 312,400 Glassdoor employee reviews in total.

Official Announcement Companies can create their own Twitter accounts and publish marketing information to the public. We considered the tweets from a brand's Twitter account as its *Official Announcements*, and collected 680,056 tweets all together.

6.4.3 Feature Extraction from Linguistic Footprints

For each facet of the social media data, the collection of M text documents forms a corpus $D = \{d_1, d_2, ..., d_M\}$, where a document d_j is a sequence of N words, i.e., $d_j = \{w_{1,j}, w_{2,j}, ..., w_{N,j}\}$, and $w_{i,j}$ is the i-th word in the d_j document. To extract semantic features from the text corpus, we employed Latent Dirichlet Allocation (LDA) [15], a topic modeling technique that analyzes topics in text documents. LDA assumes that any document d_j is modeled as a probability distribution θ_j over K topics, while any topic $k \in K$ is characterized by a probability distribution φ_k over vocabulary V.

Given the Dirichlet priors α and β, the probability distribution of the topic model can be described as:

$$P(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(z_{j,t}|\theta_j) P(w_{j,t}|\varphi_{z_{j,t}}), \tag{6.3}$$

where $\mathbf{W}$, $\mathbf{Z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ denote the vector version of the variables w_{ij}, z_{ij} (topic assignment for word w_{ij}), θ_j and φ_k respectively. Maximizing the above likelihood function by Bayesian inference [15] with parameters α and β, 200 topics were extracted as representative semantic features in the documents for each social media factor. Each topic is summarized by a set of representative words, which are associated with relevant texts.

6.4.4 Brand Personality Computation

As described in Section 6.4.1, a personality trait can be modeled as a linear combination of semantic features. In order to map semantic features to personality traits, we built a

regression model for brand personality prediction based on observed personality scales. The observation data was obtained from a survey of 10,950 valid responses on the 219 brands, where 3,060 participants rated how descriptive the personality traits were of a brand using a 7-point scale. Each brand was treated as one observation and each personality trait was regarded as a dependent variable, so that 219 observations and 42 dependent variables were included. Seven feature descriptors from the distribution of documents over each feature were used as predictors (predictor variables): mean, 5th to 95th percentile, variance, skew, kurtosis, minimum, and maximum. Therefore, the input of the prediction model has totally 4,200 predictors (200 features $\times$ 7 descriptors $\times$ 3 factors). Since the number of predictors exceeds the number of observations with a high collinearity between predictors, we used Lasso regularized regression [138]:

$$L(\gamma_1, ..., \gamma_P) = \arg\min \|\mathbf{y} - \sum_{j=1}^{P} \gamma_j \mathbf{x}_j\|^2, \text{subject to} \sum_{j=1}^{P} |\gamma_j| \leq t, \tag{6.4}$$

with P the number of predictors, $\mathbf{x}_j$ the predictors, $\mathbf{y}$ the dependent variables, γ_j the regression coefficients, and t the Lasso parameter. Lasso is able to seek for a sparse solution by constraining the coefficients of weak and correlated predictor variables to zero [138].

The relative importance of the three social media factors is evaluated based on the associated coefficients after the Lasso regression fits the observation data. The weight of a factor with respect to a trait is calculated by summing the standardized coefficients of the predictors used to predict the corresponding dependent variable:

$$W(P_k) = \sum_{j=1}^{P_k} \gamma_j \frac{SD(\mathbf{x}_j)}{SD(\mathbf{y})}, P_k \subset P, \tag{6.5}$$

where P_k is the set of predictors that belong to the k-th social media factor. The average weight of a factor on a broad trait is calculated by averaging the weights over all its subtraits.

The prediction model was evaluated with the observation data being ground truth. 10-fold cross validation was used to assess the model accuracy. The evaluation showed our model achieved predicted R^2 as high as 0.67, and mean absolute error as low as 0.01 on a $0-1$ scale. A comprehensive report of the evaluation can be found in [165].

After predicting the 42 personality traits, the values of the five broad traits were aggregated accordingly based on the trait hierarchy (as shown in Figure 6.2). Thus, for each brand, our computational approach generates a comprehensive multi-level model (as described in Section 6.4.1) that consists of the derived personality traits linked with the linguistic evidence from social media data. From now on, we denote the output of brand personality computation as *personality profiles* for further visual analytics.

6.5 System Design

In this section, we discuss the design rationales for our system, SocialBrands, and provide an overview of the system architecture.

6.5.1 Design Rationale

The domain tasks suggest that a visual analytics system is urgently needed to enable brand managers to understand and analyze public perceptions of brand personality. We identified a set of design goals as follows to address these analysis tasks based on our initial investigation with the domain experts.

G1. **Computational Brand Personality Analytics.** To save the time and efforts of our domain users in repeatedly conducting surveys for assessing perceived brand personality, it is crucial to provide a computational workflow that automatically derives personality scales from public social media data (T1). More importantly, this design could effectively improve their daily work performance in other analysis tasks (T2-T5).

G2. **Intuitive Self-Explanatory Metaphor.** Brand personality is intrinsically complex and multi-faceted. A visual metaphor capable of intuitively conveying an integrated sense of multi-dimensional brand personality is preferred by our domain users for the analysis tasks (T2-T4). To facilitate analytical reasoning and accountable explanation, the visual metaphor should be able to show how the perceived brand personality can be derived from the driving factors in social media with the support of visual evidence and related details. Varying contributions of social media factors on different personality traits should be visually highlighted to enhance the understanding of association between social media factors and brand perceptions.

G3. **Visual Analysis of Brand Comparison.** To support the analysis tasks (T2-T5), a visualization system should offer comparative analysis of two brands in terms of how similar or different their brand personality is and what the driving factors are to lead such similarity or difference. Since our domain users are concerned with examining and comparing personality profiles of related brands, the desired visual metaphor should help them easily distinguish one from another for comparative analysis.

G4. **Visual Aggregation and Summarization.** In marketing and brand analysis, brand managers often classify brands into different market segments that share common needs and interests in the marketplace. The domain experts are concerned with identifying and analyzing different market segments (T5). A visual overview of perceived personalities of a set of brands is desired by our domain users. The overview should present a contextual understanding of the marketplace of many brands by providing summarization information such as the distribution of brands over different trait dimensions and the clusters of brands.

G5. **Interactive Visual Exploration.** A visualization system that enables brand managers to interact with the data directly and see the perception results immediately is always

preferred by the domain experts to complete the described tasks (T2-T5). A visual analytical workflow with rich interactions should be provided to allow brand managers to gain insights into a brand's personality profile and its association with multi-faceted social media factors, and view related topics and discussions to completely understand public perceptions of brands on social media. They should also be allowed to flexibly explore perceived personalities of different brands in a collection.

6.5.2 System Overview

With these design goals in mind, we developed the SocialBrands system. Figure 6.4 illustrates the system architecture and the analytical workflow. The system consists of three modules: (1) the data collection and preprocessing module, (2) the brand personality modeling module, and (3) the interactive visualization module. The data collection and preprocessing module collects and stores the online social media text documents such as tweets, online reviews, and user profiles to account for three facets of linguistic footprints that drive perceptions of brand personality on social media. Tokenization and stemming are performed to preprocess the social media text documents into words as the basic linguistic evidence. The brand personality modeling module extracts topics from the words as representative semantic features in the documents for each factor, and then predicts brand personality through Lasso regression, where personality traits are regarded as dependent variables and topics as predictor variables. Furthermore, given the multi-dimensional hierarchical personality profiles produced by the brand personality model, the interactive visualization module transforms them into multiple interactive views to represent brand personality at various finer granularities for different analysis tasks. For example, an individual view and a comparative view are designed to illustrate the association between

brand perceptions and social media driving factors with linguistic evidence. An overview is designed to illustrate the visual aggregation and summarization of perceived personality of all the user-provided brands.

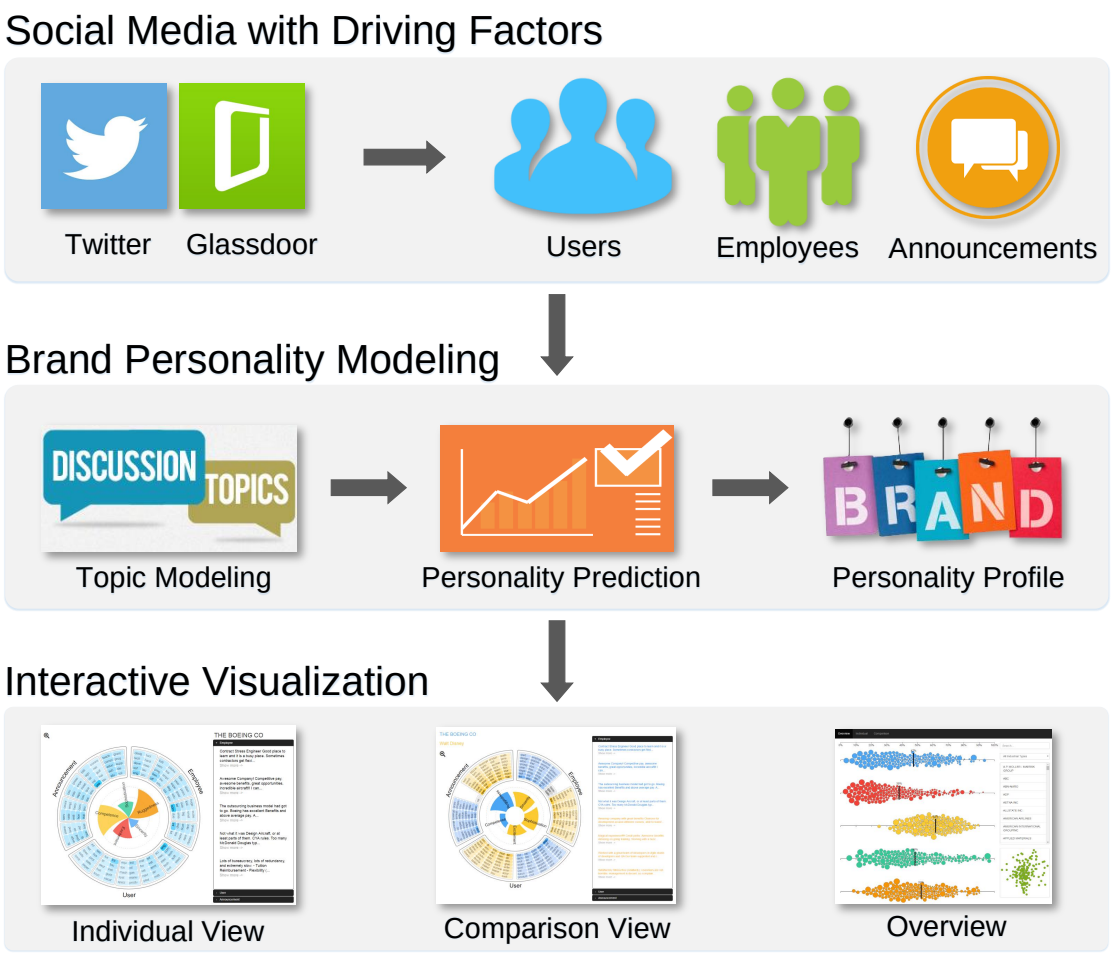


Figure 6.4: Overview of the SocialBrands system and analytical workflow.

6.6 Visualizing Brand Perceptions with SocialBrands

Following the design goals (G2-G5), we develop an interactive visualization module for SocialBrands, which consists of three major views: an Individual View, a Comparative View and an Overview. Both the **Individual View** and the **Comparative View** include a *BrandWheel* visualization (Figure 6.1(a,b,c)) with an additional document panel that shows related social media documents regarding each driving factor. The **Overview** (Figure 6.1(d)) consists of (1) multiple *BrandSediments* for visual aggregation of multi-dimensional brand personalities, (2) a list view of brands with search and filter widgets, and (3) a multidimensional scaling (MDS) view that embeds brands based on their overall personalities [83].

In the following sections, we first introduce a wheel-inspired visual metaphor for visual summarization of the perceived brand personality and visual explanation of the relation of brand personality to the driving factors from social media (Section 6.6.1), and then present a sedimentation-based visual metaphor for visual aggregation of perceived personalities in a brand collection (Section 6.6.2). Finally, we describe the interactions (Section 6.6.3) and provide a usage scenario (Section 6.6.4).

6.6.1 BrandWheels: Visual Summarization and Explanation of Brand Personality

To convey an integrated and organic sense of multi-dimensional brand personality with accountable visual explanation from linguistic perspective (G2), we design the *BrandWheel* visualization to illustrate a brand's personality profile and its association with social media factors (Figure 6.6(c)). The wheel metaphor, inspired by both classic Radial Space Filling (RSF) technique [126] and the beauty of Goethe's color wheel [85], has several benefits

compared with a traditional bar chart: the wheel metaphor with radial sectors conveys a representative sense of a brand's personality, and lower memory cost with the metaphor from real life [9]; it is also aesthetically more appealing and engaging than a bar chart. We describe each of these components below.

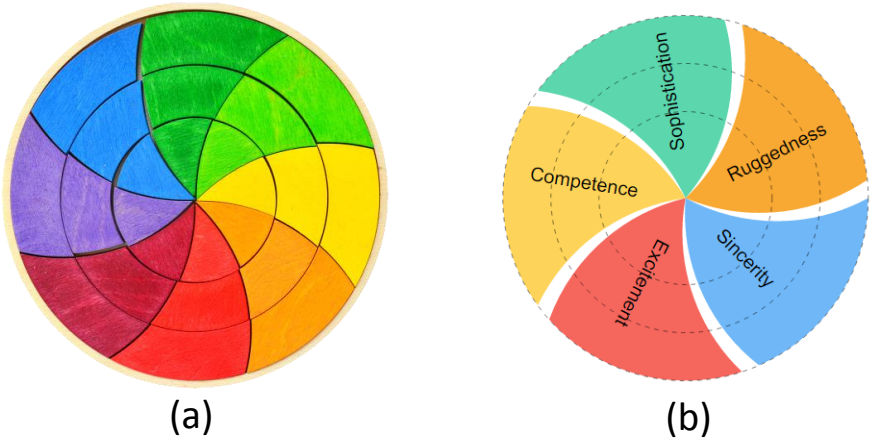


Figure 6.5: The wheel-based Metaphor in BrandWheel. (a) A jigsaw of Goethe's color wheel [85]. (b) A visual metaphor of personality wheel inspired from (a).

Personality Sectors Personality sectors form the inner component of a BrandWheel. In analogy to the Goethe's color wheel (Figure 6.5(a)), each personality sector represents one personality trait, and multiple sectors in an RSF layout illustrate a brand's personality profile with a set of personality traits (Figure 6.5(b)). Following the design of Goethe's color wheel, the *trait type* is mapped to the color of a personality sector and the radius R_t of a personality sector encodes the *trait value*: the higher a trait value is, the farther a sector spreads out (Figure 6.7(a)). Concentric circles are drawn in a dashed and grey manner

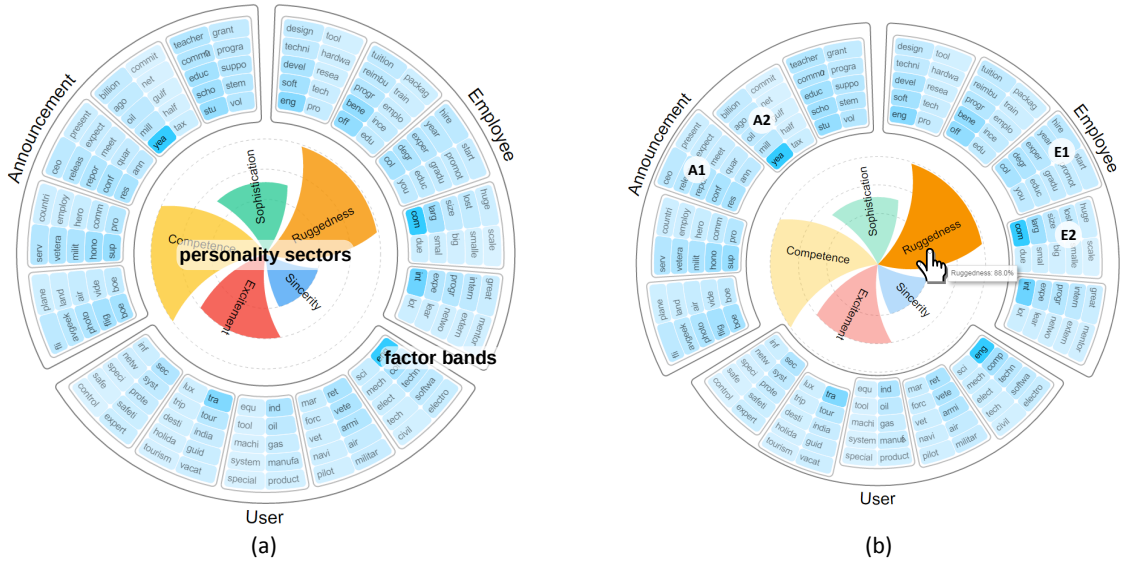


Figure 6.6: The visualization design of BrandWheel. (a) The visual metaphor of BrandWheel, composed of personality sectors (illustrating personality traits) and factor bands (showing linguistic evidence from social media factors). (b) The BrandWheel visualization with a focus on the *Ruggedness* trait: the User-factor-band gets closer to the view center than the other two, indicating a higher contribution in perceiving Ruggedness; the closeness of topic blocks (e.g., A1, A2, E1, E2) towards the center also indicates their specific relevance to Ruggedness.

to show the value range of traits, and serve as reference scales for visual alignment (Figure 6.7(a)); meanwhile, a Tooltip will pop up to show the exact trait value when the mouse is hovering on a sector. To balance aesthetics and functionality, curved wheel sectors are used to achieve the visual appearance of Goethe's color wheel. We note that the perception of a trait's value does not significantly depend on whether the sector is curved or not as a trait's value is encoded by a sector's radius.

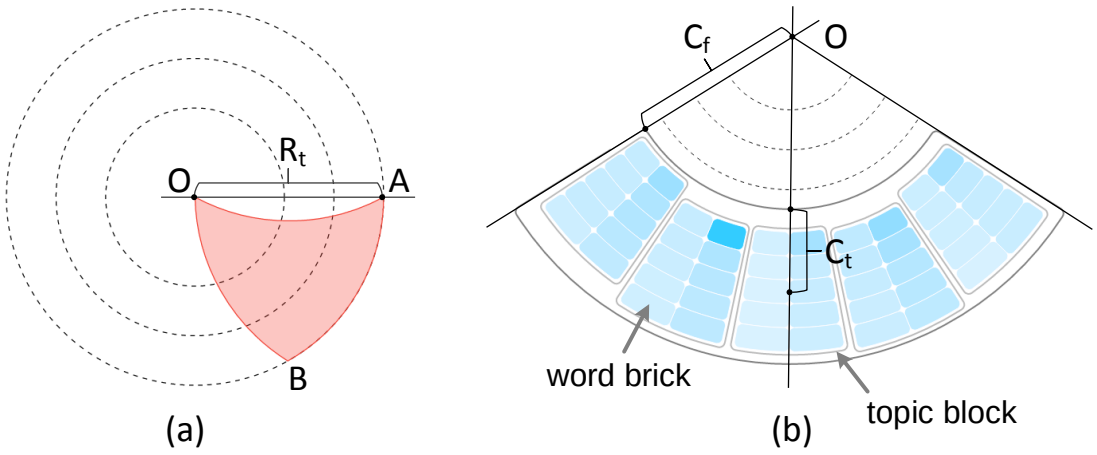


Figure 6.7: The visual encodings of BrandWheel. (a) The radius R_t of a personality sector shows a trait value. (b) The closeness C_f of a factor band (with respect to the view center) depicts the weight of the corresponding social media factor; the closeness C_t of a topic block (with respect to the inner boundary of the belonging factor band) depicts the weight of the corresponding topic regarding a particular trait; the color intensity I_w of a word brick encodes the frequency of the word in the underlying linguistic documents.

Factor Bands Surrounding the personality sectors, factor bands are shown in the outer component of a BrandWheel in an RSF layout. To support visual explanation of the influence of each driving social media factor on different personality traits, the closeness C_f of a factor band with respect to the view center (shown in Figure 6.7(b)) indicates the weight of the corresponding driving factor (defined in equation (6.5)): the closer a factor band is to the center, the more contribution it has to predict a certain trait. As three factor bands are initially aligned in an RSF layout, changes of the factors' weights when focusing on a specific trait are prominent due to the effectiveness of the visual channel *aligned spatial position* [109], as illustrated in Figure 6.6(b). To help users understand and verify the perceived personality traits, factor bands include the linguistic evidence used to derived the

traits from the social media factors in our data model (Figure 6.3). The most popular topics (i.e., salient semantic features in the evidence network) are visualized in *topic blocks*, which are tiled by *word bricks* of the representative words in an RSF layout (Figure 6.7(b)). Similar to the impact of factors on different traits, the extracted topics may also have various contributions to predict different traits (i.e., the regression coefficient γ_j of a topic in equation (6.4) can vary in different regressions regarding different dependent variables). To encode such impact of topics in the visualization, the closeness C_t of a topic block (with respect to the inner boundary of the associated factor band) is used to show the weight of the corresponding topic with respect to a trait of interest, as shown in Figure 6.7(b). To show the popularity of words in the related topics, the color intensity I_w of a word brick encodes the frequency of the word in the underlying linguistic documents. When selecting a word brick of interest, a text panel along with the BrandWheel visualization shows relevant texts with highlighted topic words to help users further understand the linguistic footprints (Figure 6.8(b)).

Comparative BrandWheels To support visual comparison of two brand personality profiles (G3), we extend the BrandWheel design and create *Comparative BrandWheel* by highlighting the similarities and differences of two brands' personality profiles with representative topical words (Figure 6.9). In this comparison view, personality sectors explicitly encode the absolute value differences of the traits of two brands (the center is made hollow to differentiate such comparison sectors from the individual sectors in a standard BrandWheel). The sectors are radially reordered by their actual value differences for visual classification of the personality traits into winning and losing traits when comparing a brand to another. Two distinct colors are mapped to the two underlying brands. In particular,

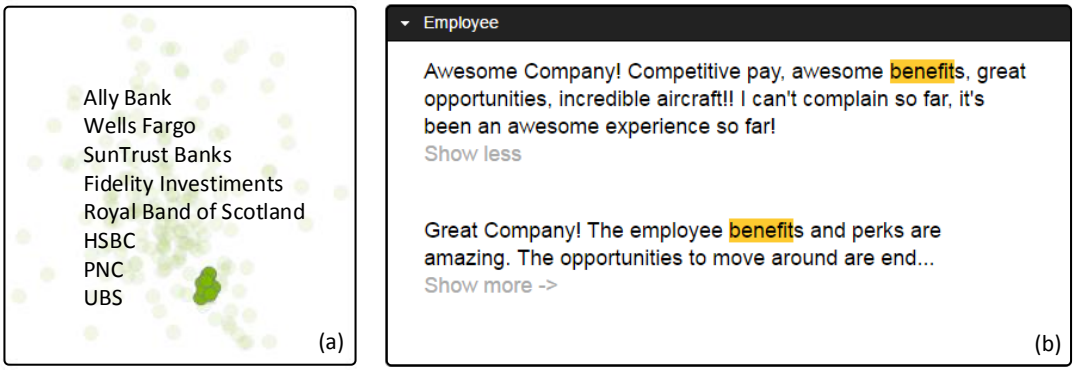


Figure 6.8: Visualization components. (a) A group of related brands identified as financial firms in the MDS view. (b) A text panel along with the BrandWheel visualization, showing relevant text documents with highlighted topic words (e.g., "benefit") for each social media factor that contribute to a trait of interest.

the color of a personality sector is mapped to the brand that has a higher value in the corresponding personality trait, which supports easy identification of the winning and losing traits of a brand. To visualize the similarities and differences of the topical words of two brands (denoted as A and B), the word bricks of two brands in each factor band are grouped into three topic blocks based on their logical relations: words only related to A, words shared by A and B, and words only related to B. In this way, the visualization of topic blocks in a factor band is reminiscent of a *Venn diagram* [154] that shows logical relations of two sets, as shown in Figure 6.9. To support interactive exploration, the individual personality sectors of a brand can be retrieved when hovering on a brand, and the topic blocks associated with the focus brand will be highlighted accordingly.

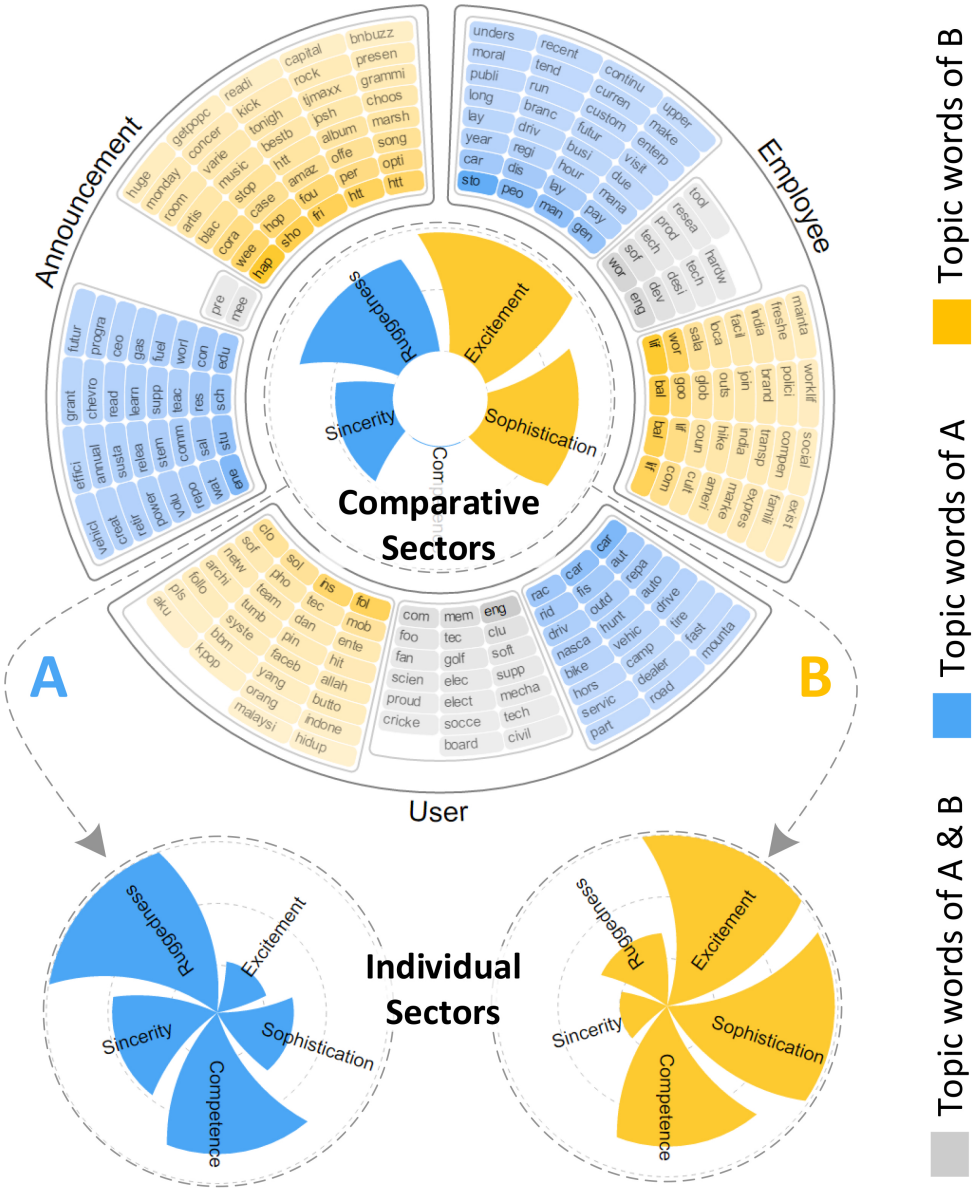


Figure 6.9: The visualization design of Comparative BrandWheel. Personality sectors explicitly encode the absolute value differences of traits of two brands. The color of a personality sector is mapped to the brand that has a higher value in the corresponding personality trait (personality sectors of a brand can be retrieved when hovering on a brand). The word bricks of two brands in each factor band are grouped into three topic blocks based on their logical relations.

6.6.2 BrandSediments: Visual Aggregation of Perceived Brand Personality

To provide a context of the marketplace where a collection of brands stand over the multiple trait dimensions (G4), we design the *BrandSediments* visualization using a sedimentation metaphor [60]. This metaphor is inspired by real-world sedimentation processes where objects aggregate into strata over time. In contrast to conventional sedimentation designs for temporal data streams [60, 172], in our BrandSediments design, brands aggregate into strata over a brand personality scale. Multiple BrandSediments are drawn side by side simultaneously to illustrate visual aggregation of brands over the scales of different personality traits. Figure 6.10 shows an example of BrandSediments aggregating brand perceptions of two personality traits *Sincerity* and *Competence*.

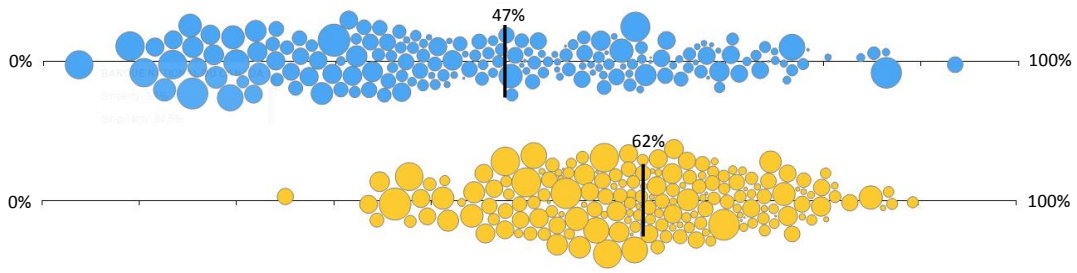


Figure 6.10: BrandSediments of the overall perceptions of brands on *Sincerity* (top) and *Competence* (bottom) in a brand collection. From an aggregated perspective, these brands are mostly perceived as competitors while sharing a varying degree of perceptions regarding sincerity .

Visual Encoding Each bubble in BrandSediments represents a brand. The colors of bubbles reflect their personality traits, which are consistent with the colors used in the BrandWheel visualization. The size of a bubble encodes *brand singularity*, which measures the degree of focus and single-mindedness of a brand's overall personality scales (brands of high singularity tend to leave an enduring impression among consumers [100]). As the BrandSediment metaphor is designed for visual aggregation of brands over a personality scale, the horizontal position of a bubble depicts the trait value of the corresponding brand, and brands of similar trait values are aggregated vertically to form a personality-based market segment within a trait value interval. To achieve such visual sedimentation, we employ a force-directed method with a horizontal constraint and collision detection to densely pack bubbles while reducing overlaps along the vertical direction. A vertical bar highlights the median value of the overall trait value distribution as a visual reference point.

6.6.3 Interactions

SocialBrands supports various types of user interactions (demonstrated in the accompanying video).

Semantic zooming allows users to drill down into a broad trait to view its subtraits in both the BrandWheel and the BrandSediments visualization. By default, five broad traits were shown to emphasize an aggregated sense of perceived personality.

Highlighting, searching and filtering allows users to interact with brands of interest for exploratory analysis. For instance, hovering over a bubble in the BrandSediments (or a brand in the list view) automatically highlights all the bubbles representing the selected brand in the BrandSediments. Users can also look for a particular brand in the search box

(Figure 6.1(d-2)), or select a group of brands that fall into a specific industrial category; the other brands will be filtered out in the BrandSediments.

Brushing and linking enables users to interactively specify a region in BrandSediments (or in the MDS view) to select a group of brands that have similar personality trait(s), and view how they are aggregated in other BrandSediments. The list view is also linked with the other views for showing the brands in a brushed region.

6.6.4 Usage Scenario

To illustrate the capability of *SocialBrands*, we describe a scenario of investigating the perceived brand personalities using our system. Suppose Emma is a brand manager whose job is to assess brand perceptions on social media. She used SocialBrands for this purpose. Her data have been processed and analyzed by the SocialBrands system for a collection of brands of her interest.

To start with, Emma viewed the aggregation of brand perceptions in the SocialBrands Overview (Figure **??**(d)). She observed a cluster of brands that share similar overall brand personalities in the MDS view (Figure 6.8(a)), which were identified as financial companies such as Ally Bank, Wells Fargo and HSBC. She then focused on the *Competence* trait in the BrandSediments visualization. After selecting a small subset of highly competent brands by brushing the BrandSediments, she found that they represent high-tech companies such as Boeing, Apple, and Cisco. She then drilled down into the *Competence* trait to see the underlying personality traits, and observed that most brands are extremely *corporate*.

To investigate why people perceived a particular brand to be competent, Emma selected the brand "Boeing" to view its BrandWheel in the Individual View (Figure 6.6(c)). She found that Boeing's employees frequently mentioned words such as "benefits", "offer" and

"products", as revealed by the topic blocks in the Employee-factor-band. For instance, one relevant employee review on Glassdoor said that "Awesome Company! Competitive pay, awesome benefits, great opportunities, incredible aircraft!" (Figure 6.8(b)). To gain further understanding of the perceived personality of Boeing, Emma selected the *Ruggedness* trait. From the displacement of factor brands (from Figure 6.6(c) to Figure 6.6(d)), she learned that the *User Imagery* factor affects more on perceiving Boeing as *rugged* as the User-factor-band moved closer to the view center. Furthermore, from the displacement of topic blocks, she observed that the topic popularity remains roughly the same in the User-factor-band, while those in the other two showed prominent changes. For instance, topic A2 (with words "million", "tax", etc.) becomes more relevant to *Ruggedness* than A1 (with words "result", "report", etc.) in the Announcement-factor-band, while the topic E2 (with words "company", ""large", etc.) is more related to *Ruggedness* than E1 (with words "cool", "you", etc.) in the Employee-factor-band.

As a follow-up study, Emma selected another brand "Disney" to compare its perceived personality with "Boeing" in the Comparative View. As shown in Figure 6.9, Boeing (A) and Disney (B) have distinct perceptions in the *Ruggedness* and *Sincerity* traits — Boeing is perceived as much more rugged than Disney, while Disney is perceived as far more sincere than Boeing. To explain such differences, she resorted to the data evidence in the factor bands. She learned that the social media discussions about these two brands are very different, as they share few common topic words. For example, Disney's employees frequently mentioned words such as "free", "amazing", and "member", while Boeing's employees preferred to discuss "intern", "engineering", and "offer".

6.7 Evaluation

SocialBrands is the very first visual analytic system for brand personality analysis. Thus, we adopted an exploratory study with a think-aloud protocol that helps us understand how domain users naturally use such a new system, and collects rich information when users freely express their own thoughts during exploration. We evaluated SocialBrands through in-depth interviews and surveys with marketing brand managers. The interviews and surveys helped us obtain detailed qualitative feedback from the domain users.

6.7.1 Participants and Procedure

Participants for both the interviews and the surveys were 12 brand managers from different marketing departments such as sales, consumer research, and advertising in a large international IT company. Their management positions ranged from managers and directors to vice president, and their experiences ranged from 4-24 years. A majority of them ($n = 10$) have worked at multiple companies for brand management. We denote the participating managers as M1-M12. The data set used for interviews contains 181 brands of leading companies in various industrial sectors including electronics, banking, entertainment, health care and so on.

Interviews lasted one hour and were audio recorded and transcribed respectively. Interviews were split into three stages: a pre-interview (10 minutes), a system walkthrough (40 minutes) and a post-interview (10 minutes). In the pre-interview, we asked the participants about their experience and training in marketing and branding. Then we asked the participants how familiar they were with the concept of brand personality, and what they thought of the personality of their company and their clients. In the system walkthrough,

participants were introduced to the system and requested to explore it openly with a think-aloud protocol. In the post-interview, we asked the participants what they learned about the brands, and how this system worked to meet their expectations. Then we asked the participants when and how they might use this type of system in their marketing or branding work. Finally, we asked the participants what parts of the interface were particularly useful to them, and what parts were confusing or otherwise need improvement. Afterwards, participants took a post-interview survey, which contains 12 seven-point Likert-scale questions about the usefulness and usability of SocialBrands (Table 6.1).

6.7.2 Results

Our evaluation captured a continuous exploration process of using such a new system. Overall the brand managers responded positively to our SocialBrands system and found it useful for exploring, assessing and interpreting perceived brand personality on social media (see survey results in Table 6.1). Below we report the study results of the SocialBrands system in several key aspects.

Assessing Brand Perceptions All participants found the SocialBrands system highly useful in assessing brand personality both in general (Q1, μ=6.4, σ=0.52) and individually (Q2, μ=6.6, σ=0.52), which are among the most helpful features based on the ratings. Before using the system, participants were only able to make vague and general evaluation about how people perceived their brands. They thought perceptions of a brand were very difficult to describe in an comprehensive and accurate manner. For example, M1 stated that *"there are probably a lot of instances that people may have an idea of a company, but they cannot put fingers on what exactly that is"*. In contrast, after using SocialBrands, participants were able to have a quick grasp of the overall personality pattern for a brand, and

Table 6.1: Usefulness and usability of SocialBrands. The average ratings (μ) with standard deviations (σ) from brand managers. 7 means "strongly agree" with a statement, 1 means "strongly disagree" with it, and 4 indicates "neutral".

Question	μ	σ
Q1. Overall, assessed the personality of brands	6.4	0.52
Q2. Assessed a brand's personality profile	6.6	0.52
Q3. Identified gaps between the inferred and the intended personality of a brand	5.4	1.06
Q4. Understood relationships between a brand's personality and contributing factors	5.3	1.28
Q5. Identified gaps between brands	6.5	0.53
Q6. Identified patterns of brands based on personality	5.9	0.99
Q7. Easy to learn and use SocialBrands	6.0	0.93
Q8. The *Personality Sectors* design conveyed an integrated sense of brand personality	6.1	0.64
Q9. The *Factor Bands* design was helpful for connecting the personality to contributing factors	5.1	1.36
Q10Easy to identify winning/losing traits of a brand in the Comparative View	6.0	0.76
Q11Easy to find relevant brands in the Overview	6.6	0.52
Q12Willing to use a system similar to SocialBrands for brand management in the future	6.1	0.64

concretely discuss their understanding of perceived brand personality in a more thorough and confident fashion, as the personality sectors of BrandWheel conveyed a compact and organic sense of brand personality (Q8, μ=6.1, σ=0.64). M1 commented that *"the circular layout [of BrandWheel] is intuitive in terms of quickly identifying the best and the worst traits of a brand"*.

Visual Explanation of Brand Perceptions A majority of the participants highly appreciated the idea of providing linguistic evidence for visual explanation of brand perceptions. They were interested to see the actual data that behind the perceptions of a brand, and were excited to discuss different topics emerging in the BrandWheel. For instance, M3 said *"this [BrandWheel] is really useful, powerful and believable — as a marketer, it certainly helps me to explore how my brand is represented across these key social media platforms; otherwise, I just have to believe whatever the personality is"*. Furthermore, many participants confirmed the effectiveness of the interactive movements of factor bands in revealing the changes of contributions of social media factors when focusing on different traits. M9 remarked that he was *"able to see which factor is most associated with [a trait], and what are the differences of the associations among factors"*. Nevertheless, a few participants mentioned that it was still difficult to reason a factor's effects based on linguistic footprints. M12 commented that *"the importance of the texts is not quantified or prioritized"*. Consequently, the variance of survey ratings on Q9 is relatively high (σ=1.36), though the average rating is still on the positive side (μ=5.1).

Identification of Perception Gaps Many participants expressed that our system assisted them in identifying personality gaps between perceived and intended personality of a brand (Q3, μ=5.4, σ=1.06). Specifically, participants were able to validate their marketing strategies by examining dominate traits that they wanted to emphasize, and identifying weakly-perceived traits that need to be targeted or improved further. M2 said that *"we are doing better on several key personality attributes that I had told you initially; maybe a little worse on sincerity than I thought, but higher on competence"*. With a better understanding

of these gaps in personality perceptions, brand managers are more likely to form target-oriented marketing strategies to improve brand perceptions on specific trait dimensions.

Comparative Analysis of Brand Perceptions The Comparative BrandWheel visualization helped participants effectively compare personalities of selected brands, and identify gaps between them (Q5, μ=6.5, σ=0.53). Many participants reported that comparing brands with similar personalities can help them discover potential collaborators. M1 said that *"[Comparative BrandWheel] is really great to compare two brands; focusing on 'accomplished', 'cool' and 'innovated' can leverage what you know about similar brands to find complementary clients"*. Furthermore, Comparative BrandWheels supported participants to learn from winning brands (Q10, μ=6.0, σ=0.76). M3 commented that *"[Comparative BrandWheel] tells who may be strong or weak on particular personality traits, helping me understand what our brand needs to do"*. In particular, the visualization of topic blocks provided specific examples to learn from winning brands. M9 remarked that *"the topic words outside support me to see the differences that might be driving the differences on the inside [traits]"*.

Perception-based Market Segmentation Most participants enjoyed the BrandSediments visualization (Q11, μ=6.6, σ=0.52), enabling a contextual understanding of the marketplace by showing where a brand stands with regard to other representative companies. This contextual understanding facilitated participants to gain insights into the distribution of brands over personality traits. M4 commented that *"exploring the brands around us gives me a better idea of what was meant by 'original'; these brands help me figure out how to interpret that trait"*. Moreover, participants used our system to easily and systematically identify market segments that share similar personality patterns (Q6, μ=5.9, σ=0.99).

Such perception-based market segments can be used in competitor analysis. M11 remarked that *"we could use it [the BrandSediments visualization] to identify upcoming competitors competitors based on brand personality"*.

Generalizability of Visual Design Participants were very much impressed with the design of BrandWheels and BrandSediments, and believed that they were generalizable for various marketing purposes. In particular, one participant suggested that interfaces like BrandWheels and BrandSediments can be used by chief marketing officers (CMOs) and chief financial officers (CFO) to shape the company's understanding of a particular product, sales strategy or marketing idea: *"if we come up with five market segments that we think are important to CMOs and CFOs, this [BrandWheel] could be the way to visualize, communicate, and explore segments"*.

Overall System Usability We learned that many participants thought SocialBrands was intuitive, easy to learn and use (Q7, μ=6.0, σ=0.93). The metaphors of BrandWheel and BrandSediments were considered intuitive, explanatory and engaging. M8 commented that *"it is pretty easy to navigate [BrandSediments]; this [BrandWheel] is laid out pretty well"*. They reported that the majority of features in the system were useful and practical. M6 stated that *"we need more such social tools to help us understand what social conversations are around us and other brands and how we compare in personality"*.

6.8 Discussion

Prior to our work, brand managers mostly used surveys to collect descriptive ratings on brand personality scales to assess brand personality. Our evaluation of SocialBrands demonstrated the usefulness and usability of this very first visual analytic system for brand

personality analysis in the hands of brand managers. We discuss the design implications for developing such visual analytic tools in marketing industry.

First of all, SocialBrands is a successful example of applying marketing theories (e.g., brand personality) in developing a domain-specific application. Future system designers are encouraged to make use of domain literature and theoretical frameworks that are powerful to explain and guide the practices in the domain for users. This also suggests that the design of visual analytics tools is desirable to root in a theoretical foundation, amplify human's understanding of the domain tasks with interactive visualization, and also loop human's knowledge into the theoretical framework [135].

Second, SocialBrands encodes multi-level data evidence in the visualization (i.e., factor bands) for domain users to interpret and explain the analytics results (e.g., personality traits). Future visualization researchers are advocated to design intuitive metaphors that enable domain users to easily relate analytics results to their evidence at multiple levels and facilitate their analytical reasoning. The domain experts can collect and synthesize such evidence to either confirm or discard the hypotheses in their minds, and the intuitive visual metaphors of evidence serve as an important resort to build the loop in the visual analytics process [75].

Third, SocialBrands exposes the transparency of the computational model to the domain users for them to understand how the model works. The most challenging task in our system design was to help users understand how perceived personality was derived from different social media factors. While we implicitly encodes the contributions of factors through interaction (i.e., displacements of factor bands), alternative designs such as explicit encoding through visual links [127] may also be considered to address this open problem. Furthermore, with a deeper understanding of how the model works, the domain users can

interact with the analysis results and give the feedback to the system (e.g., increasing the weight of a factor of interest when deriving perceived personality). Such feedback from domain users can be further fed into the model to refine the analysis results and close the human-in-loop visual analytics. This brings in new challenges for designing visual analytics tools and also is the future direction of this work.

There are several interesting directions for generalizing and extending our current system. First, our system can be easily integrated with existing social marketing tools [111, 116] to bring a new perspective about brand performance. Second, as a strong brand can be leveraged across multiple sub-brands to achieve specialized marketing purposes, our system can be used to analyze perceptions of sub-brands to help brand managers understand the roles of sub-brands in a brand's architecture. Third, we plan to conduct graphical perception studies to investigate the functionality and engagement of curved charts against bar charts by measuring interpretation accuracy and long-term recall. Finally, we would like to supplement SocialBrands with historical marketing abilities to analyze brand perceptions over time.

6.9 Summary

In this chapter, we present SocialBrands, a novel visual analysis tool for brand managers to understand public perceptions of brands on social media. Based on branding theories and advanced social computing approaches, we compute the perceived brand personality with linguistic features extracted from three driving factors on social media – user imagery, employee imagery and official announcement, and quantify the associations between brand perceptions and social media factors. The computational results are integrated with new

visualization designs for visual aggregation, summarization and explanation of brand perceptions. Through in-depth interviews and surveys with brand managers, we find that brand managers can gain new insights into brand perceptions in five ways: (a) assessing and analyzing brand perceptions on social media, (b) identifying gaps between the intended and the perceived brand personality, (c) understanding multiple factors that drive brand personality on social media, (d) learning from successful brands through visual comparison, and (e) discovering new perception-based market segments.

Chapter 7: Conclusions

In this final chapter, we first summarize the contributions of this research, with a focus on the approaches used to address the research questions posed at the beginning of this dissertation. Then we discuss the perspectives of this research and outline promising future research directions. Finally, we conclude this dissertation with broader remarks of this research.

7.1 Contributions

The fundamental goal of this dissertation is to investigate critical aspects of multidimensional data visualization and comparative analytics in assisting users in visual exploration of multidimensional data for knowledge discovery and sense-making. Specifically, we address the following two high-level research questions:

Question 1: How can we design multidimensional visualizations to enable effective visual summarization of multi-faceted data characteristics?

Question 2: How can we enhance the power of multidimensional visualizations with comparative analytics to allow visual identification and explanation of complex data relationships?

Based on the theoretical foundations of visualization and visual analytics, we approach the questions with various research methodologies such as creative design methods, quantitative studies, and qualitative studies. This dissertation has contributed to the design and understanding of comparative visualization and analysis methods in helping users explore multidimensional data for discovering knowledge, making decisions and communicating insights within multi-faceted information spaces. Below we summarize the main contributions.

7.1.1 Comparative Multidimensional Visualization Techniques

- *A new comparative visualization — CorrelatedMultiples*, creating spatially coherent juxtaposition of multidimensional data items. Incorporating similarity into the spatial layout, viewers can better assess the overall qualitative characteristics of the multidimensional data. We modeled the relationships among data items as a similarity graph, and embedded the items in the layout so that proximity reflects similarity. CorrelatedMultiples reduces viewers' visual search space when searching for similar items near the target, dissimilar items far from it, or anomalies at the periphery of the layout.

- *A compact visualization — TileMatrix*, coupling the side-by-side juxtaposition, back-to-back juxtaposition and complementary juxtaposition to display a large number of adjacency matrices. We described an application scenario of TileMatrix for visualizing multi-faceted, time-varying weighted networks. The repeated patterns and differences in the relationships of data items can be identified and understood from the TileMatrix representation, in both data and temporal domains.

- *A composite visualization — AssociationGraph* that visualizes confident scalar-level associations in multiple variables and their derived attributes. A multi-partite graph visualization is designed to capture the relationships of scalars from different variables, which is linked with a juxtaposed visualization of multiple rank-based parallel coordinates plots for exploring salient scalars with confident associations.

- *A multi-faceted visualization — BrandWheel* that conveys an integrated sense of multidimensional brand personality with visual evidence and related details. Composed of multidimensional personality sectors and factor bands, BrandWheel enables visual explanation of how the perceived brand personality can be derived from the driving factors in social media, and visually highlights varying contributions of social media factors on different personality traits.

- *A comparative visualization — BrandSediment* that enables a contextual understanding of the marketplace of many brands. Using a sedimentation metaphor inspired by real-world sedimentation processes, the juxtaposition of multiple BrandSediments facilitates perception-based market segmentation through visual summarization of the distribution of brands over different personality dimensions and the clusters of brands.

7.1.2 Multidimensional Data Analysis Methods

- *An optimization algorithm — Constrained Multi-Dimensional Scaling (CMDS)* that produces an optimal proximity graph layout within a given display space. We showed that this optimization approach is computationally efficient, visually stable and space-efficient through numerical and visual comparisons with other state-of-the-art methods. In addition, it is also easy to implement the algorithm with an iterative solver.

- *An association analysis method* that models the directional interactions between values of different variables as information flows based on association rules, and computes the informativeness and uniqueness of values based on information flows. We proposed the Multi-Scalar Informativeness-Uniqueness (MSIU) algorithm, based on an information propagation model in social networks, to quantify informativeness and uniqueness from scalar-level associations.

- *A computational analysis approach* that is designed to assess brand personality from three driving factors of user imagery, employee imagery and official announcement on social media, and construct a multi-faceted evidence data explaining the association between brand personality and its driving factors.

7.1.3 Design Implications and Guidelines

- *Design implications for choosing representations and juxtapositions of adjacency matrix visualizations.* Triangular representation does not hamper graphical perception of adjacency matrices. Symmetric juxtaposition rather than translational juxtaposition should be preferred for detecting changes of structures and patterns. Complementary juxtaposition is beneficial for optimizing utilization of display space.

- *Guidelines for visual exploration of scalar-level associations of different variables in the multivariate data space.* We extended the visualization mantra "Overview first, zoom and filter, then details-on-demand" [124] to visual exploration of informative and unique scalar values and their confident associations in multivariate data: *overview first* by the PCP and PAGraph views, *zoom and filter* to narrow down the

focus to a particular scalar of interest through interactive brushing, and view *details-on-demand* of the spatial relationship of the selected scalar and its associated scalars of other variables in the spatial view.

- *Design implications for developing visual analytic tools in marketing industry.* Future system designers are encouraged to make use of domain literature and theoretical frameworks that are powerful to explain and guide the practices in the domain for users. Future visualization researchers are advocated to design intuitive metaphors that enable domain users to easily relate analytics results to their evidence at multiple levels and facilitate their analytical reasoning.

7.1.4 Generalizable Visual Exploration Frameworks

- *A visual analytic framework* with multiple interactive views for data scientists to explore the scalar values of interest with confident associations in the multivariate spatial domain. Representing directional interactions between scalar values as information flows using a probabilistic association graph, our framework makes a novel use of an information propagation model from the social network literature to analyze scalar values in the multivariate domain, and brings a novel perspective to visual exploration of multivariate data sets.

- *A visual analytic tool — SocialBrands* for brand managers to assess and analyze public perceptions of brands on social media. Our evaluation of SocialBrands demonstrated the usefulness and usability of this very first visual analytic system for brand personality analysis in the hands of brand managers. SocialBrands is a successful example of applying marketing theories in developing a domain-specific application. It encodes multi-level data evidence in the visualization for domain users to interpret

and explain the analytics results. It also exposes the transparency of the computational model to the domain users for them to understand how the model works. With a deeper understanding of how the model works, the domain users can interact with the analysis results and give the feedback to the system.

7.2 Prototypes

The research prototypes developed in this dissertation were mostly implemented as web-based systems, which are manageable and compatible across platforms. In this section, we describe the implementation details of the prototypes.

- We implemented the CMDS algorithm in HTML5 and Javascript to create the prototype of CorrelatedMultiples. Because we did not find a suitable red-black tree library in Javascript for implementing the scan-line algorithm, we substituted with a naïve $O(|V|^2)$ node collision detection algorithm. This is reasonable because a typical computer display of about 1000×1000 pixels can at most fit about 400 objects if we allow each object to be about 50×50 pixels, so $|V|$ is not that large. For larger $|V|$, a log-linear collision detection algorithm such as a scan-line algorithm would still be desirable.

- The TileMatrix visualization prototype was implemented using D3.js [16], a JavaScript library that binds data to DOM elements in HTLM5. Each tile in TileMatrix is implemented by a *matrix* module, which can be configured in one of four possible triangular orientations. Multiple triangular matrices were bound to SVGs in HTML5 to achieve the effect of tiling.

- The association-guided exploration framework was implemented using HTLM5, Javascript and Python. The Python-based computation component includes calculation of 1D histograms for each variable, calculation of 2D joint histograms for each pair of variables, construction of probabilistic association graph, and computation of informativeness and uniqueness values. The radial graph visualization was adapted from the *arc* module and the *chord* layout in D3.js. The rank-based PCP visualization was implemented using the modules of *line*, *axis* and *brush* in D3.js.

- The SocialBrands system was implemented using HTLM5, D3.js and Java. The Java-based computation component consists of data collection, tokenization, stemming, LDA and Lasso regression. Twitter API and Glassdoor API were used to collect social media data from three contributing factors of brand personality. Amazon MTurk was used to collect observation data for brand personality prediction. The web-based visualization component includes several visualization modules: the *sector* module implements the basic component of a personality sector; the *wheel* module is made up of several sectors to create a personality wheel; the *band* module implements a factor band, which is made up of topic blocks and word bricks; the *ring* module assembles multiple factor bands to form a composite visualization; the *bubbleChart* module implements the BrandSediments visualization for one personality trait, and the *multiples* module assembles multiple BrandSediments in an aligned layout.

7.3 Future Directions

Individual future directions on each of the topics covered in this dissertation have been described in the corresponding chapters. In this section we discuss the open research problems that remain and how the efforts in this dissertation can contribute.

Supporting Multi-faceted Viewing of Data. The visualization techniques proposed in this dissertation apply to multidimensional data. In addition to multiple dimensions, data sets collected in real world may come from multiple sources. For example, people may have different social networks on different social media platforms such as Twitter, Facebook and LinkedIn. Each network offers one way of looking at the social interactions among people. *How to provide a multi-faceted viewing of data with many dimensions as well as from many different sources? What would be the effective data models and representations? What transformations are needed to obtain those data models and representations?*

Generalizing Designs of Hybrid Comparative Visualizations. In this dissertation, we focused on investigating alternative designs of juxtaposition in creating effective comparative visualizations. As suggested by Gleicher et al. [42], the proposed juxtaposition designs can be combined with other comparative visualization designs such as superimposition and explicit encoding, which poses interesting challenges to hybrid comparative visualization design. *Is it possible to provide a general framework for creating and recommending alternative hybrid comparative visualizations? What are the design challenges that need to be addressed?*

Developing Practical Visual Analytic Tools. The techniques described in this system can be integrated with existing visual analytic tools to bring a new perspective on comparative analytics of multidimensional data. With a deeper understanding of a visual analytic tool, the domain users can interact with the analysis results through visualizations and provide the feedback to the system, which can be used to further refine the analysis results. *How to make these tools both usable and useful in the hands of domain users? How to*

adapt the visualization techniques to address domain users' requirements and the specific analytic tasks?

7.4 Closing Remarks

Big data is not merely the size of data, but rather that their fine-grained nature permits inference and decisions at the level of single individuals [23]. Data with high dimensionality and complexity has far exceeded our human ability for comprehension without powerful tools. This dissertation has investigated problems and presented techniques for supporting effective visual exploration of multidimensional data with comparative analytics. We hope this dissertation can significantly influence and inspire further research in the design and understanding of comparative visualizations and analysis techniques for multidimensional data.

Bibliography

[1] Jennifer L Aaker. Dimensions of brand personality. *Journal of marketing research*, pages 347–356, 1997.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD*, volume 22, pages 207–216, 1993.

[3] Hiroshi Akiba, Kwan-Liu Ma, Jacqueline H Chen, and Evatt R Hawkes. Visualizing multivariate volume data from turbulent combustion simulations. *Computing in Science and Engineering*, 9(2):76–83, 2007.

[4] Basak Alper, Benjamin Bach, Nathalie Henry Riche, Tobias Isenberg, and Jean-Daniel Fekete. Weighted graph comparison techniques for brain connectivity analysis. In *CHI*, 2013.

[5] Daniel Archambault, Helen Purchase, and Bruno Pinaud. Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):539–552, April 2011.

[6] Benjamin Bach, Emmanuel Pietriga, and Jean-Daniel Fekete. Visualizing dynamic networks with matrix cubes. In *CHI*, 2014.

[7] Stefano Baldassi, Nicola Megna, and David C Burr. Visual clutter causes high-magnitude errors. *PLoS Biol*, 4(3):e56, 2006.

[8] Michelle Q Wang Baldonado, Allison Woodruff, and Allan Kuchinsky. Guidelines for using multiple views in information visualization. In *Advanced Visual Interfaces*, pages 110–119, 2000.

[9] Scott Bateman, Regan L Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2573–2582. ACM, 2010.

[10] Richard A Becker and William S Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[11] Michael Behrisch, James Davey, Fabian Fischer, Olivier Thonnard, Tobias Schreck, Daniel Keim, and Jörn Kohlhammer. Visual analysis of sets of heterogeneous matrices using projection-based distance functions and semantic zoom. *Computer Graphics Forum*, 33(3):411–420, 2014.

[12] Jacques Bertin. *Graphics and graphic information processing*. Walter de Gruyter, 1981.

[13] Jacques Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[14] Ayan Biswas, Soumya Dutta, Han-Wei Shen, and Jonathan Woodring. An information-aware framework for exploring multivariate data sets. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2683–2692, 2013.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of machine Learning research*, 3:993–1022, 2003.

[16] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[17] 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.

[18] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[19] Stuart K Card, Allen Newell, and Thomas P Moran. The psychology of human-computer interaction. 1983.

[20] MST Carpendale. Considering visual variables as a basis for information visualisation. 2003.

[21] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 1984.

[22] Scott Cohen and L Guibasm. The earth mover's distance under transformation sets. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1076–1083. IEEE, 1999.

[23] N Council. Frontiers in massive data analysis, 2013.

[24] Weiwei Cui, Yingcai Wu, Shixia Liu, Furu Wei, Michelle X Zhou, and Huamin Qu. Context preserving dynamic word cloud visualization. In *Pacific Visualization Symposium (PacificVis), 2010 IEEE*, pages 121–128. IEEE, 2010.

[25] Etienne Cuvelier and Marie-Aude Aufaure. A buzz and e-reputation monitoring tool for twitter based on galois lattices. In *International Conference on Conceptual Structures*, pages 91–103. Springer, 2011.

[26] Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer, 3rd edition, April 2008.

[27] Kasper Dinkla, Michel A Westenberg, and Jarke J van Wijk. Compressed adjacency matrices: untangling gene regulatory networks. *TVCG*, 18(12):2457–2466, 2012.

[28] Marian Dork, Daniel Gruen, Carey Williamson, and Sheelagh Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138, 2010.

[29] T. Dwyer, K. Marriott, and P. J. Stuckey. Fast node overlap removal. In *Proc. 13th Intl. Symp. Graph Drawing (GD '05)*, volume 3843 of *LNCS*, pages 153–164. Springer, 2006.

[30] Andrew SC Ehrenberg, Mark D Uncles, and Gerald J Goodhardt. Understanding brand performance measures: using dirichlet benchmarks. *Journal of Business Research*, 57(12):1307–1325, 2004.

[31] Geoffrey Ellis and Alan Dix. A taxonomy of clutter reduction for information visualisation. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1216–1223, 2007.

[32] D. Eppstein, M. van Kreveld, B. Speckmann, and F. Staals. Improved grid map layout by point set matching. In *IEEE Pacific Visualization Symposium*, 2013.

[33] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[34] Imola K Fodor. A survey of dimension reduction techniques, 2002.

[35] Johannes Fuchs, Fabian Fischer, Florian Mansmann, Enrico Bertini, and Petra Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *CHI*, pages 3237–3246. ACM, 2013.

[36] Raphael Fuchs and Helwig Hauser. Visualization of multi-variate scientific data. *Computer Graphics Forum*, pages 1670–1690, 2009.

[37] Karl Ruben Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[38] Emden R. Gansner and Yifan Hu. Efficient node overlap removal using a proximity stress model. In *Graph Drawing*, pages 206–217. Springer-Verlag, Berlin, Heidelberg, 2009.

[39] Emden R. Gansner, Yehuda Koren, and Stephen North. Graph drawing by stress majorization. In *International Conference on Graph Drawing*, 2004.

[40] Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. On the readability of graphs using node-link and matrix-based representations: A controlled experiment and statistical analysis. *Information Visualization*, July 2005.

[41] Glassdoor. http://www.glassdoor.com, 2016.

[42] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[43] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D Hansen, and Jonathan C Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[44] GlpkJS. A GNU linear programming kit for Javascript, http://hgourvest.github.com/glpk.js/.

[45] E. Gomez-Nieto, W. Casaca, L.G. Nonato, and G. Taubin. Mixed integer optimization for layout arrangement. In *Conference on Graphics, Patterns and Images*, pages 115–122, 2013.

[46] Erick Gomez-Nieto, Frizzi San Roman, Paulo Pagliosa, Wallace Casaca, E Helou, M Ferreira de Oliveira, and L Nonato. Similarity preserving snippet-based visualization of web search results. *Visualization and Computer Graphics, IEEE Transactions on*, 2013.

[47] Luke J Gosink, John C Anderson, E Wes Bethel, and Kenneth I Joy. Variable interactions in query-driven visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1400–1407, 2007.

[48] Liang Gou, Michelle X Zhou, and Huahai Yang. KnowMe and ShareMe: Understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 955–964. ACM, 2014.

[49] Haleh Hagh-Shenas, Sunghee Kim, Victoria Interrante, and Chris Healey. Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1270–1277, 2007.

[50] Derek L Hansen, Dana Rotman, Elizabeth Bonsignore, Nataa Milic-Frayling, Eduarda Mendes Rodrigues, Marc Smith, and Ben Shneiderman. Do You Know the

Way to SNA?: A process model for analyzing and visualizing social media network data. In *International Conference on Social Informatics*, pages 304–313. IEEE, 2012.

[51] Steve Haroz and David Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2402–2410, 2012.

[52] Rom Harré and Roger Lamb. *The dictionary of personality and social psychology*. Blackwell Oxford, 1986.

[53] John A Hartigan. Printer graphics for clustering. *Journal of Statistical Computation and Simulation*, 4(3):187–213, 1975.

[54] Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *CHI*, 2009.

[55] N. Henry, J. Fekete, and M.J. McGuffin. Nodetrix: a hybrid visualization of social networks. *TVCG*, 13(6):1302–1309, Nov 2007.

[56] Nathalie Henry and Jean-Daniel Fekete. Matrixexplorer: a dual-representation system to explore social networks. *TVCG*, 12(5):677–684, 2006.

[57] Nathalie Henry and Jean-Daniel Fekete. Matlink: Enhanced matrix visualization for analyzing social networks. In *Human-computer Interaction*, 2007.

[58] Patrick Hoffman, Georges Grinstein, Kenneth Marx, Ivo Grosse, and Eugene Stanley. Dna visual and analytic data mining. In *Visualization'97., Proceedings*, pages 437–441. IEEE, 1997.

[59] Xiaodi Huang, Wei Lai, ASM Sajeev, and Junbin Gao. A new algorithm for removing node overlapping in graph visualization. *Information Sciences*, 177(14):2821–2844, 2007.

[60] Samuel Huron, Romain Vuillemot, and Jean-Daniel Fekete. Visual sedimentation. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2446–2455, 2013.

[61] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[62] Takayuki Itoh, Chris Muelder, Kwan-Liu Ma, and Jun Sese. A hybrid space-filling and force-directed layout method for visualizing multiple-category graphs. In *IEEE Pacific Visualization Symposium*, 2009.

[63] Heike Janicke, Michael Bottinger, and Gerik Scheuermann. Brushing of attribute clouds for the visualization of multivariate data. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1459–1466, 2008.

[64] Heike Janicke, Alexander Wiebel, Gerik Scheuermann, and Wolfgang Kollmann. Multifield visualization using local statistical complexity. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1384–1391, 2007.

[65] Waqas Javed and Niklas Elmqvist. Exploring the design space of composite visualization. In *Pacific Visualization Symposium*, pages 1–8. IEEE, 2012.

[66] Waqas Javed, Bryan McDonnel, and Niklas Elmqvist. Graphical perception of multiple time series. *TVCG*, 16(6), 2010.

[67] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[68] M Chris Jones and Robin Sibson. What is projection pursuit? *Journal of the Royal Statistical Society. Series A (General)*, pages 1–37, 1987.

[69] Fabien Jourdan and Guy Melancon. Multiscale hybrid mds. In *International Conference on Information Visualisation*, 2004.

[70] Sepandar D Kamvar and Jonathan Harris. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 117–126. ACM, 2011.

[71] Eser Kandogan. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of the IEEE Information Visualization Symposium*, volume 650, page 22. Citeseer, 2000.

[72] Jean-Noel Kapferer. *The new strategic brand management: Advanced insights and strategic thinking*. Kogan page publishers, 2012.

[73] Johannes Kehrer and Helwig Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *Visualization and Computer Graphics, IEEE Transactions on*, 19(3):495–513, 2013.

[74] Johannes Kehrer, Harald Piringer, Wolfgang Berger, and M Eduard Groller. A model for structure-based comparison of many categories in small-multiple displays. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2287–2296, 2013.

[75] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer, 2008.

[76] Daniel Keim et al. Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):1–8, 2002.

[77] René Keller, Claudia M. Eckert, and P. John Clarkson. Matrices or node-link diagrams: Which visual representation is better for visualising connectivity models? *Information Visualization*, 5(1):62–76, March 2006.

[78] D Brett King and Michael Wertheimer. *Max Wertheimer and gestalt theory*. Transaction Publishers, 2005.

[79] T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. *Self-Organizing Maps*. Springer-Verlag, 2001.

[80] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[81] Thomas Kraft, Derek Xiaoyu Wang, Jeffrey Delawder, Wenwen Dou, Yu Li, and William Ribarsky. Less After-the-Fact: Investigative visual analysis of events from streaming twitter. In *IEEE Symposium on Large-Scale Data Analysis and Visualization*, pages 95–103. Citeseer, 2013.

[82] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[83] Joseph B Kruskal and Myron Wish. *Multidimensional scaling*, volume 11. Sage, 1978.

[84] I. Kruskal. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer Verlag, 1997.

[85] Kylie. How we montessori, http://www.howwemontessori.com/how-we-montessori/crafts/, 2016.

[86] Heidi Lam, Tamara Munzner, and Robert Kincaid. Overview use in multiple visual information resolution interfaces. *TVCG*, 13(6):1278–1285, 2007.

[87] Roger Th AJ Leenders. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24(1):21–47, 2002.

[88] Marc Levoy. Display of surfaces from volume data. *Computer Graphics and Applications, IEEE*, 8(3):29–37, 1988.

[89] Xiaotong Liu, Yifan Hu, Stephen North, and Han-Wei Shen. Compactmap: A mental map preserving visual interface for streaming text data. In *Big Data, 2013 IEEE International Conference on*, pages 48–55. IEEE, 2013.

[90] Xiaotong Liu, Yifan Hu, Stephen North, and Han-Wei Shen. Correlatedmultiples: Spatially coherent small multiples with constrained multi-dimensional scaling. In *Computer Graphics Forum*. Wiley Online Library, 2015.

[91] Xiaotong Liu, Srinivasan Parthasarathy, Han-Wei Shen, and Yifan Hu. Galaxyexplorer: Influence-driven visual exploration of context-specific social media interactions. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 215–218, 2015.

[92] Xiaotong Liu and Han-Wei Shen. The effects of representation and juxtaposition on graphical perception of matrix visualization. In *Proceedings of the 33rd annual ACM conference on Human factors in computing systems*, pages 269–278. ACM, 2015.

[93] Xiaotong Liu and Han-Wei Shen. Association analysis for visual exploration of multivariate scientific data sets. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):955–964, 2016.

[94] Xiaotong Liu, Han-Wei Shen, and Yifan Hu. Supporting multifaceted viewing of word clouds with focus+context display. *Information Visualization*, May 2014.

[95] Xiaotong Liu, Anbang Xu, Liang Gou, Haibin Liu, Rama Akkiraju, and Han-Wei Shen. Socialbrands: Visual analysis of public perceptions of brands on social media. In *IEEE Conference on Visual Analytics Science and Technology*, 2016.

[96] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM siggraph computer graphics*, pages 163–169. ACM, 1987.

[97] Alan M MacEachren. Visualizing uncertain information. *Cartographic Perspectives*, (13):10–19, 1992.

[98] Alan M MacEachren. *How maps work: representation, visualization, and design*. Guilford Press, 2004.

[99] Jalal Mahmud, Geli Fei, Anbang Xu, Aditya Pal, and Michelle Zhou. Predicting attitude and actions of twitter users. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 2–6. ACM, 2016.

[100] Lucia Malär, Bettina Nyffenegger, Harley Krohmer, and Wayne D Hoyer. Implementing an intended brand personality: a dyadic perspective. *Journal of the Academy of Marketing Science*, 40(5):728–744, 2012.

[101] Muhammad Muddassir Malik, Christoph Heinzl, and M Eduard Groeller. Comparative visualization for parameter studies of dataset series. *Visualization and Computer Graphics, IEEE Transactions on*, 16(5):829–840, 2010.

[102] Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. TwitInfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–236. ACM, 2011.

[103] A.R. Martin and M.O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Visualization, 1995. Visualization '95. Proceedings., IEEE Conference on*, pages 271–, 1995.

[104] Pierre Martineau. The personality of the retail store. 1958.

[105] James R Miller. Attribute blocks: Visualizing multiple continuously defined attributes. *Computer Graphics and Applications, IEEE*, 27(3):57–69, 2007.

[106] Charles Joseph Minard. *Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813*. Graphics Press., 1869.

[107] Joel L Morrison. A theoretical framework for cartographic generalization with the emphasis on the process of symbolization. *International Yearbook of Cartography*, 14(1974):115–27, 1974.

[108] Christopher Mueller, Benjamin Martin, and Andrew Lumsdaine. Interpreting large visual similarity matrices. In *International Asia-Pacific Symposium on Visualization*, pages 149–152, 2007.

[109] Tamara Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[110] Suthambhara Nagaraj, Vijay Natarajan, and Ravi S Nanjundiah. A gradient-based comparison measure for visual analysis of multifield data. In *Computer Graphics Forum*, volume 30, pages 1101–1110. Wiley Online Library, 2011.

[111] Oktopost. www.oktopost.com, 2015.

[112] Hans-Georg Pagendarm and Frits H Post. *Comparative visualization-approaches and examples*. Delft University of Technology, 1995.

[113] Brian T Parker. A comparison of brand personality and brand user-imagery congruence. *Journal of Consumer Marketing*, 26(3):175–184, 2009.

[114] Charles Perin, Frédéric Vernier, and Jean-Daniel Fekete. Interactive horizon graphs: improving the compact visualization of multiple time series. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3217–3226. ACM, 2013.

[115] Roberto Dantas Pinho, Maria Cristina Oliveira, and Alneu Andrade Lopes. An incremental space to visualize dynamic data sets. *Multimedia Tools Appl.*, 50(3):533–562, December 2010.

[116] Rignite. www.rignite.com, 2015.

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

[118] George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko. Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1325–1332, November 2008.

[119] K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. Evaluating a visualisation of image similarity as a tool for image browsing. In *IEEE Symposium on Information Visualization*, 1999.

[120] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. Influence and passivity in social media. In *Machine learning and knowledge discovery in databases*, pages 18–33. Springer, 2011.

[121] Takafumi Saito, Hiroko Nakamura Miyamura, Mitsuyoshi Yamamoto, Hiroki Saito, Yuka Hoshiya, and Takumi Kaseda. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 173–180. IEEE, 2005.

[122] Natascha Sauber, Holger Theisel, and Hans-Peter Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5), September 2006.

[123] Claudia Schmid and Hans Hinterberger. Comparative multivariate visualization across conceptually different graphic displays. In *Scientific and Statistical Database Management, 1994. Proceedings., Seventh International Working Conference on*, pages 42–51. IEEE, 1994.

[124] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343. IEEE, 1996.

[125] Daniel J Simons and Daniel T Levin. Change blindness. *Trends in cognitive sciences*, 1(7):261–267, 1997.

[126] John Stasko and Eugene Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization*, pages 57–65. IEEE, 2000.

[127] Markus Steinberger, Manuela Waldner, Marc Streit, Alexander Lex, and Dieter Schmalstieg. Context-preserving visual links. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2249–2258, 2011.

[128] H. Strobelt, M. Spicker, A. Stoffel, D. Keim, and O. Deussen. Rolled-out wordles: A heuristic method for overlap removal of 2d data representatives. *Computer Graphics Forum*, 31(3pt3):1135–1144, 2012.

[129] Grant Strong and Minglun Gong. Data organization and visualization using self-sorting map. In *Graphics Interface*, pages 199–206, 2011.

[130] Grant Strong, Rune Jensen, Minglun Gong, and AnneC. Elster. Organizing visual data in structured layout by maximizing similarity-proximity correlation. In *Advances in Visual Computing*. Springer, 2013.

[131] Julius Tsu-li Su. *An electron force field for simulating large scale excited electron dynamics*. PhD thesis, California Institute of Technology, 2007.

[132] Jinshan Tang, R.M. Rangayyan, Jun Xu, I El Naqa, and Yongyi Yang. Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances. *IEEE Transactions on Information Technology in Biomedicine*, 13(2):236–251, March 2009.

[133] Aditya Tatu, F Maas, Ines Farber, Enrico Bertini, Tobias Schreck, Thomas Seidl, and Daniel Keim. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *IEEE Conference on Visual Analytics Science and Technology*, pages 63–72. IEEE, 2012.

[134] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[135] James J Thomas. *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005.

[136] James J Thomas and Kristin A Cook. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Laboratory (PNNL), Richland, WA (US), 2005.

[137] Norman JW Thrower. *Maps and civilization: cartography in culture and society*. University of Chicago Press, 2008.

[138] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[139] New York Times. Drought's footprint: http://www.nytimes.com/interactive/2012/07/20/us/drought-footprint.html.

[140] New York Times. Mapping migration in the united states: http://www.nytimes.com/2014/08/16/upshot/mapping-migration-in-the-united-states-since-1900.html.

[141] USA TODAY. Online nba statistics. http://www.usatoday.com/sports/nba/statistics/.

[142] Christian Tominski, Camilla Forsell, and Jimmy Johansson. Interaction support for visual comparison inspired by natural behavior. *TVCG*, 18(12):2719–2728, 2012.

[143] E. R. Tufte. *Envisioning Information*. Graphic Press, 1990.

[144] Edward R Tufte. Envisioning information. *Optometry & Vision Science*, 68(4):322–324, 1991.

[145] Edward R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press LLC, 1997.

[146] Edward R Tufte and PR Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

[147] John W Tukey. Exploratory data analysis. 1977.

[148] Cagatay Turkay, Peter Filzmoser, and Helwig Hauser. Brushing dimensions-a dual visual analysis model for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2591–2599, 2011.

[149] Twitter. https://support.twitter.com/articles/166337-the-twitterglossary, 2016.

[150] Christopher W Tyler. *Human symmetry perception and its computational analysis*. Psychology Press, 2003.

[151] Pravin Vaidya. Geometry helps in matching. In *ACM Symposium on Theory of Computing*, pages 422–425, 1988.

[152] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[153] Jarke J Van Wijk. The value of visualization. In *Visualization, 2005. VIS 05. IEEE*, pages 79–86. IEEE, 2005.

[154] John Venn. On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(59):1–18, 1880.

[155] Fernanda Viégas, Martin Wattenberg, Jack Hebert, Geoffrey Borggaard, Alison Cichowlas, Jonathan Feinberg, Jon Orwant, and Christopher Wren. Google+ ripples: A native visualization of information flow. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1389–1398. ACM, 2013.

[156] Chaoli Wang, Hongfeng Yu, Ray W Grout, Kwan-Liu Ma, and Jacqueline H Chen. Analyzing information transfer in time-varying multivariate data. In *Pacific Visualization Symposium*, pages 99–106. IEEE, 2011.

[157] Junpeng Wang, Xiaotong Liu, Han-Wei Shen, and Guang Lin. Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots. *IEEE Transactions on Visualization and Computer Graphics*, 2016.

[158] Yang Wang, Liang Gou, Anbang Xu, Michelle X Zhou, Huahai Yang, and Hernan Badenes. VeilMe: An interactive visualization tool for privacy configuration of using personality traits. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 817–826. ACM, 2015.

[159] Colin Ware. *Visual thinking: For design*. Morgan Kaufmann, 2010.

[160] Colin Ware. *Information visualization: perception for design*. Elsevier, 2012.

[161] Kazuho Watanabe, Hsiang-Yun Wu, Yusuke Niibe, Shigeo Takahashi, and Issei Fujishiro. Analyzing information transfer in time-varying multivariate data. In *Pacific Visualization Symposium*. IEEE, 2015.

[162] Pak Chung Wong and R. Daniel Bergeron. 30 years of multidimensional multivariate visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques*, pages 3–33, Washington, DC, USA, 1997. IEEE Computer Society.

[163] Pak Chung Wong and R Daniel Bergeron. Multivariate visualization using metric scaling. In *Visualization'97., Proceedings*, pages 111–118. IEEE, 1997.

[164] Jo Wood and Jason Dykes. Spatially ordered treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1348–1355, 2008.

[165] Anbang Xu, Haibin Liu, Liang Gou, Rama Akkiraju, Jalal Mahmud, Vibha Sinha, Yuheng Hu, and Mu Qiao. Predicting perceived brand personality with social media. In *The International AAAI Conference on Web and Social Media*, 2016.

[166] Panpan Xu, Yingcai Wu, Enxun Wei, Tai-Quan Peng, Shixia Liu, Jonathan JH Zhu, and Huamin Qu. Visual analysis of topic competition on social media. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2012–2021, 2013.

[167] Di Yang, Elke A Rundensteiner, and Matthew O Ward. Analysis guided visual exploration of multivariate data. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 83–90. IEEE, 2007.

[168] Tal Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373, 2010.

[169] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[170] Xiaoru Yuan, Peihong Guo, He Xiao, Hong Zhou, and Huamin Qu. Scattering points in parallel coordinates. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1001–1008, 2009.

[171] Xiaoru Yuan, Donghao Ren, Zuchao Wang, and Cong Guo. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2625–2633, 2013.

[172] Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu-Ru Lin, and Christopher M Collins. #FluxFlow: Visual analysis of anomalous information spreading on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1773–1782, 2014.

[173] Jian Zhao, Liang Gou, Fei Wang, and Michelle Zhou. Pearl: an interactive visual analytic tool for understanding personal emotion style derived from social media. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 203–212. IEEE, 2014.