THREE-COMPONENT VISUAL SUMMARY: A DESIGN TO SUPPORT CASUAL EXPERTS IN MAKING DATA-DRIVEN DECISIONS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Calvin Yau

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. David S. Ebert, Chair

School of Electrical and Computer Engineering

Dr. Alexander J. Quinn

School of Electrical and Computer Engineering

Dr. Edward J. Delp

School of Electrical and Computer Engineering

Dr. Niklas Elmqvist

College of Information Studies, University of Maryland

Approved by:

Dr. Dimitrios Peroulis

Head of the School Graduate Program

To my parents.

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Dr. David S. Ebert, for his guidance and support throughout my Ph.D. journey. He taught me how to do research, helped me refine my ideas, connected me with valuable resources, provided for me financially through the research assistantship, and was patient with me as I inevitably made mistakes along the way.

I would also like to thank my committee members Dr. Niklas Elmqvist, Dr. Alex Quinn, and Dr. Edward Delp, for their helpful feedback and advice.

Next, I would like to thank my VACCINE colleagues for their help and their friendships throughout my time in the lab. Their company made this supposedly lonely journey a lot more bearable.

Additionally, I would like to thank my roommate Lester and my friends from Clear River Church, Graduate InterVarsity Christian Fellowship, One A Chord A Capella Group, and Purdue University Choir for all the joy and support they brought into my life throughout the seven years.

Likewise, I would like to thank my parents for valuing education and for encouraging me when I faced hardships because of this path I chose, my brother Paul for providing a friendship like no other even when we were physically far apart, and my fiancée Jane for always lifting me up during the most difficult times and for making me a better person in every way.

Lastly, I would like to thank my savior Jesus Christ for being the source of my hope and strength.

Year 2020 marks a significant milestone in my life, and I would not have been able to make it here by my own strength, so thank you.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Yau, Calvin Ph.D., Purdue University, May 2020. Three-Component Visual Summary: A Design to Support Casual Experts in Making Data-Driven Decisions. Major Professor: David S. Ebert.

Recent advancements in data-collecting technologies have posed new opportunities and challenges to making data-driven decisions. While visual analytics can be a powerful tool for exploring large datasets and extracting relevant insights to support data-driven decisions, many decision-makers lack the time or the technical expertise to utilize visual analytics effectively. It is more common for data analysts to explore data through visual analytics and report their findings to the decision-makers. However, the communication gap between data analysts and decision-makers limits the decision-maker's ability to make optimal data-driven decisions. I present a Three-Component Visual Summary to allow accurate and efficient extraction of insights relevant to the decisions and provide context to validate the insights retrieved. The Three-Component Visual Summary design creates visual summaries by combining visual representations of representative data, analytical highlights, and the data envelope. This design incorporates a high-level summary, the relevant analytical insights, and detailed explorations into one coherent visual representation which addresses the potential training gaps and limited available time for visual analytics. I demonstrate how the design can be applied to four major data types commonly used in commercial visual analytics tools. The evaluations prove the design allows more accurate and efficient knowledge retrieval and a more comprehensive understanding of the data and of the insights generated, making it more accessible to decision-makers that are casual experts. Finally, I summarize the insights gained from the design process and the feedback received, and provide a list of recommendations for designing a Three-Component Visual Summary.

1. INTRODUCTION

With the convenience of modern data-collecting technology, we are generating data faster than ever before. As a result, interest in exploring this increasing volume of data has expanded beyond trained data analysts to the average consumer. More everyday technology users, with expertise in different technical disciplines, are interested in using data gathered about their subjects of interest to help make data-driven decisions. For example, YouTube artists can use YouTube Analytics [1] to understand viewer demographics and view duration to improve their content; farmers bury sensors under their farmland to monitor the moisture and temperature of the soil at different depths to adjust their irrigation to achieve optimal yield [2]; and small companies embed cameras on their advertisement boards to examine prospective customers at different locations in order to adjust their advertisement strategies [3]. However, even with data collected and tools developed to navigate through the data, there remain gaps that prevent many such audiences from effectively and efficiently generating insights for decision making.

1.1 Challenges in Exploring and Analyzing Data for the Casual-Expert Decision-Makers

It has been observed from the law enforcement and first responder fields that decision-makers are interested in making data-driven decisions, but often lack formal training and expertise in data analysis or visual analytics. This section explores the main challenges that prevent such decision-makers from utilizing their data effectively.

First, there is too much data to make sense of without having technical support. Examining data one entry at a time in a spreadsheet or a folder is not only inefficient but also

[1] https://www.youtube.com/analytics?o=U

[2] http://www.vinsense.net/

[3] https://www.admobilize.com/

ineffective in obtaining a comprehensive view of the dataset and comparing the relationships between the different data entries. Various techniques, including statistics, information visualization, and visual analytics, have been introduced to address the challenge of making sense of large datasets. Statistical metrics, such as mean, median, and mode, describe the characteristics of a large dataset in an aggregated manner. The drawback to using only statistical metrics is that they provide little insights into the individual data points and allow little to no exploration of the original dataset. Information Visualization [1] provides a more natural way to understand the stories behind the data using visual representations of data. However, even with the use of visualization techniques, computing devices generally do not have enough pixels to render and therefore communicate larger datasets properly. An interactive visual analytics system allows its audience to explore a visualized dataset at different scales to gain a comprehensive understanding [2]. This practice, however, requires dedicating time to interactive exploration and training in using visual analytics, which decision-makers do not always have.

Second, the majority of the solutions described above are intended for data analysts. Many decision-makers I worked with as a part of my thesis research are casual experts [3] - people with strong domain knowledge in their corresponding fields who possess an understanding of basic technologies but have little training in data analysis, information visualization, or visual analytics. Casual experts with domain specific knowledge can often provide explanations or suggest the best action to follow when presented with the appropriate information extracted from the data. Many of these decision-makers expressed interest in supporting their decisions with data in order to form stronger arguments and minimize the impact of potential personal bias in the matter. Unfortunately, the majority of visualization designs and visual analytics tools require sufficient knowledge of the strengths and weaknesses of different visualization designs and interactive functionalities. For example, one can obtain a more precise comparison between two different datasets by adjusting the scale of the data displayed. But having the axis starting at a different point could confuse human perceptions of how one dataset compares to the other in the overall scope.

Third, data summaries rarely emphasize significant data entries and cannot be explored backward to understand and compare the original dataset. Casual experts are often in roles where decisions need to be made within a limited amount of time. They need to be able to understand the data and its significance to the decisions they need to make efficiently and effectively, which is why a summary of the data has to be provided. However, the majority of data summarization methods focus on the aggregation of data [4] and do not emphasize on specific data entries, which may not be the best practice in the context of decision making. During the decision-making process, casual experts explore the data with a specific goal in mind, whether it is to find the company to invest in, to identify which project to cut, or to understand the strengths and weaknesses of a product. Casual experts often need to be able to identify and compare specific details contained in the raw data. Therefore, some data entries or the insights derived from them can be more important than other data entries, making traditional summarization methods insufficient for casual experts in decision making.

A common practice that has arisen in response to a lack of time and formal training in data analysis among casual experts is for trained data analysts to prepare reports and data visualizations, which are then presented to casual-expert decision-makers. This practice leads to another challenge. Decision-makers are often presented with unintentionally biased information due to the differences in training and expertise of the data analysts. While a data analyst may have the same interests and goals as the decision-maker, the decision-maker is rarely given the tools or time to examine and identify possible potential information bias. Not having the opportunity to participate in the interactive exploration process during the use of visual analytics tools also means that the decision-makers may not be able to obtain the knowledge of how the data was processed and filtered. It can be difficult for decision-makers to be certain they are making optimal data-driven decisions based solely on the information presented to them.

This research work aims to address the challenges in supporting casual experts with visual analytics and guiding casual experts to generate the insights needed to make data-driven decisions without the expertise in visual analytics. Six decision-makers were sur-

veyed to obtain a more in-depth understanding of the challenges. Based on the survey, I extracted a list of design requirements that are used to derive the solution proposed in this dissertation. Chapter 3 will provide more detail of the survey, the design requirements, and the solution proposed.

1.2 Designing Visual Summaries for Casual Experts

Based on the insights learned from the decision-maker survey, this research work proposes that it is important for a solution to consider large amounts of data, allow efficient insight generation, accommodate the skill level of casual experts, and provide context to the analysis results to support casual experts in making data-driven decisions. To satisfy these characteristics, this work draws inspiration from communication-minded visualization [5], casual information visualization [6], narrative visualization [7], and context-preserving visualization techniques [8–12].

In this work, I propose a visual summarization design called the *Three-Component Visual Summary* to support casual experts in making data-driven decisions more effectively and efficiently. Three-Component Visual Summaries reduce a large and detailed dataset into (1) *representative data* that provides a quick takeaway of the full dataset at first glance, (2) *analytical highlights/comparisons* that distinguish specific analyses of interest such as outliers, and (3) a *data envelope* that summarizes the remaining aggregated data to provide context to the analysis results. Through the three components, casual experts can quickly obtain an overview of the dataset, gain the insights required for the decision, and explore the reasoning behind the analysis results and the remaining data if so desired.

To validate the design, I applied it to the four data types most commonly used in visual analytics – numerical data, contextual data, geospatial data, and network data [13] – and evaluated the resulting products. Note that while contextual data can be any data that gives context to an entity or an event, the scope of this work focuses primarily on text data and a small amount of multi-media data.

1.3 Thesis Statement

In this work, I present the challenges in designing visual analytics that aid casual experts in accurately and efficiently understanding the relevant analysis results and context of the collected data to support data-driven decision making. I propose a Three-Component Visual Summary design as the solution to address the challenges in the volume of data, the time limitations, the potential bias in the knowledge transfer process, and the lack of training in visual analytics. The Three-Component Visual Summary hypothesizes that:

> *Domain experts with limited training in information retrieval and visual analytics can generate context-preserved insights to support decision making more accurately and efficiently through visual presentations of information by simultaneously displaying and connecting high-level overview, comparative analysis, and low-level exploration context with constrained interaction functionalities.*

This dissertation presents the design process of the Three-Component Visual Summary – the motivation, the design requirements, and the guidelines for encoding the representative data, the analytical highlights/comparisons, and the data envelope component – and discusses how this design can aid casual experts in making data-driven decisions. This dissertation then presents a set of Three-Component Visual Summary designs for numerical data, contextual data, geospatial data, and network data and evaluates the effectiveness of the designs. From the evaluations and discussions of the designs, the dissertation compiles a set of recommendations to using the Three-Component Visual Summary design.

The main contributions of this work include:

- A proposed design guideline to create a Three-Component Visual Summary utilizing the representative data, analytical highlights/comparisons, and the data envelope to allow a more effective data exploration and insight generation while reducing the dependency on interactive exploration.

- A Three-Component Visual Summary design for numerical data, contextual data, geospatial data, and network data and their corresponding evaluations in supporting data-driven decisions. The evaluations found the designs to allow a more accurate, efficient, and accessible understanding of the data and the relevant insights.
- A set of recommendations for designing and using Three-Component Visual Summaries.

1.4 Outline

This thesis has been organized into the following chapters. Chapter 2 discusses the background and related work in data communication and the four data types. Chapter 3 describes the design of the Three-Component Visual Summary. Chapter 4 presents Summarized Line Graph, a numerical Three-Component Visual Summary design that focuses on time-series data, and a quantitative study that evaluates the design against other time-series visualization techniques. Chapter 5 presents and evaluates SuccessVis, a contextual Three-Component Visual Summary design that visualizes a combination of text and multimedia data to communicate the academic impact of a research center. Chapter 6 presents and evaluates a geospatial Three-Component Visual Summary design that visualizes crime reports from Tippecanoe County, Indiana. Chapter 7 presents and evaluates a network Three-Component Visual Summary design that is built on top of the NodeTrix design [14] to visualize directed, non-weighted, ground-truth community information flow data such as Twitter following networks. Chapter 8 evaluates the four designs in conjunction and provides recommendations for applying the Three-Component Visual Summary designs. Finally, Chapter 9 provides a summary of this thesis and outlines the future work.

2. BACKGROUND AND RELATED WORK

In this chapter, I first discuss the different approaches to data abstraction/summarization. I then examine data communication approaches that may benefit casual experts [3] in understanding their data. Finally, I discuss visualization techniques for different data types and compare work relevant to the Three-Component Visual Summary designs proposed in this dissertation.

2.1 Data Abstraction

As mentioned in Chapter 1, a summary has to be provided to communicate larger datasets to the decision-makers efficiently and effectively. When summarizing scalar data, descriptive statistics (or summary statistics) is a common and well-developed approach to describe the features of a collection of information quantitatively [15]. Mean and standard deviation (sometimes with the addition of skewness and kurtosis) are commonly used to describe a snapshot of the entire dataset providing a representative value and a basis for interpreting the data through probability. However, descriptive statistics limit the ability of users to examine details of the original data sources and events that only happened to a few data sources within the group could easily be overlooked when averaged out by the other data contributors. In contrast, this work aims to allow a more detailed exploration of the subsets of the data as well as the change over time.

Sampling is also often applied to summarize large data [16]. However, when used in temporal data, sampling typically focuses on summarizing the time-axis rather than the data sources, which is different from the focus of this work. Another issue with applying data sampling to solve the research problem presented in this work, which also holds to other data abstraction techniques such as segmentation, dimension reduction, and clustering and when applied to data types different than numerical data [17], is that it treats all the data

equally and does not evaluate what information is removed in the process, whereas this work focuses on including the crucial pieces that are often overlooked.

Visual analytics has been an effective tool for exploring large volumes of data [18], and naturally, visualization designs have been developed to summarize data [19]. The majority of these designs summarize a dataset through the use of aggregation, subsampling, filtering, and projection [4]. Classic examples include bar charts [20], pie charts [21], treemaps [22], etc. Some additional techniques that allow the audience to better perceive the data through visualization designs include hierarchical aggregation [23] and aspect ratio adjustment [1, 24]. Many visual analytics systems also allow the analysts to, through some form of zoom, pan, and filter, interactively navigate through different granularities of the original dataset to examine both the high-level pattern and specific details of the dataset [25].

Sharing a similar goal to extract crucial characteristics and summarize a dataset using visual analytics, Kocherlkota et al. [26] reduce multidimensional data to its important and relevant characteristics and generate summary visualizations of the data using the extracted important characteristics. Patterns and outliers, as the important characteristics which can provide key insights for decision-making, are extracted during an interactively guided summarization process. This technique, however, has the following limitations. First, this technique relies heavily on an interactive process. Second, this technique could accidentally omit important data entries during the interactive process. Finally, this technique uses visualization techniques explicit tailored to multivariate data.

While the majority of scalar data summarization techniques can be applied to handle numerical data directly, these techniques often summarize the scalar characteristics instead of the content when applied to contextual, geospatial, or network data. For example, text-based data summarizations often focus on the frequency of unique words [27] and geospatial data summarizations often focus on the count of data points [28]; these can be useful, but may not be enough to support the casual experts in decision making as they are missing the focus points of these data types, which are the stories behind the words and the geographical locations. Therefore, preserving the context is another important aspect of data summarization to be considered in this work. Various visual analytics designs uti-

lize distortion [29, 30], multiple connected views [10–12], or overlaid annotations [8, 9] in their designs to ensure the audiences can connect back to the overview when examining the zoomed-in or filtered detail. This work follows the same approach to provide an overview and analytical highlights that are connected to the low-level details.

2.2 Data Communication

The current work falls in between the traditional use of visual analytics and that of casual information visualization [6]. Pousman and Stasko proposed the idea of casual information visualization as a complement to a more traditional information visualization domain and focus on depicting information that is more personal to casual audiences for both everyday work and non-work uses. This work shares some of the characteristics of casual information visualization, such as targeting audiences who may not be experts at reading visualizations and the design challenge of modifying the design for different users, data, and insights needed. However, this work also targets more toward specific work tasks rather than everyday tasks. Many of the decision-makers may not be trained in understanding visualization, but have domain expertise in the data presented as well as the problem to be solved.

Viegas and Wattenberg [5] introduced the concept of communication-minded visualization to support communication and collaborative analysis through visualization designs emphasizing the design of the user experiences. Inspired by their work, this work focuses on solving the specific communication gap between data analysts and their audiences through novel visualization techniques.

Segel and Heer [7] suggested design strategies for narrative visualization to tell data stories. Their work discussed the importance of balancing author-driven and reader-driven stories; my solution presented in this dissertation shares characteristics of both. Segal and Heer also suggested that storytelling of data is most effective when there is constrained interaction, which is a powerful tool to identify the important and relevant characteristics

of large, multidimensional datasets [31]. This work supports designs with constrained interaction that is limited to hover and click.

Hullman and Diakopoulos [32] presented a narrative visualization framework that uses rhetoric visualizations to tell the data stories more effectively, by using a combination of visual representations, annotations, and interactivity layers in the design. This work follows a similar strategy in the storytelling by overlaying the overview and analysis result on top of the remaining aggregated data.

IBM developed a free online data visualization tool, Many Eyes [33]. Similar to Microsoft Excel and Tableau, Many Eyes targets mass users with structured data but no programming or technical expertise, which includes casual experts. Unfortunately, the system focuses mainly on numerical data and provides its list of visualization techniques based on the data entered. Users will need to have enough knowledge of the strength and possible confusion of each technique, and will be limited to only the techniques the system can provide.

Some visualization techniques also present multiple data types to complement the storytelling. A common combination used to tell stories with more context utilizes numerical and textual data. Textual data are often overlaid on traditional numeric visualization techniques as (interactive) annotations [34, 35] to give reasons behind the numeric data behavior. This approach is more effective when displaying a comparatively smaller amount of contextual data, as it can suffer from scalability issues.

2.3 Visualizing Different Data Types

Zhang et al. classified the state-of-the-art commercial visual analytics systems into four groups by the type of data the systems support visualizing: numerical data, text/web data, geo-related data, and network data [13]. To evaluate how this work can aid the design of visual analytics in general, I apply the design to each of the data types.

Numerical data, likely the most common data type employed in visual analytics, has long been supported by different visualization techniques such as line graphs [20], scat-

terplots [36], bar charts [20], pie charts [21], etc. Many visualization techniques such as stacked graphs [37], parallel coordinates [38], box plots [39], etc. also build on these traditional techniques. Through these visualization techniques, numerical data encodes many variables of interest: financial values, weather measurements, temporal evolution, etc., and even geographical data are comprised of a combination of multiple numeric values.

Although geographical data is comprised of numerical values, geospatial visualization as a specific data type is gaining more interest in the visualization research field as more data are now geo-tagged. Geospatial visual analytics provide insights into clusters and movements and often encodes data of other types such as time [40] and text [41–43]. Such insights help first responders better analyze and predict crime reports [44], boating accidents [45], and so on.

Text-based visualization is relatively new, with the majority of the modern techniques published post-2000. Despite being a less popular visualization area (comparing to numerical-based visualization), research in text visualization has generated some power techniques such as the word cloud visualization [27]. The result from text analysis on entities, sentiments, topics, temporal evolutions, etc. can be visualized in systems such as Jigsaw [46], which is designed specifically for analyzing collections of small text documents. Visualization designs for text-based data mainly focus on displaying the trend and/or temporal evolution of entities [27, 42, 47–49], whether that is individual keywords frequently used, or topics extracted from the text collection. Through text visualization and analysis, people can better understand the trending topics on social media posts [50] or the connections between separately generated reports, etc. Asides from textual data, visualization techniques have also been developed for other contextual multimedia data such as audio and video. For example, Shibata at el. [51] visualizes the sound signals measured to diagnose faults in rotating machinery. Record the Earth [52] visualizes the locations and characteristics of sounds collected through crowd-sourcing to allow soundscape ecologists to study the correlations between the different characteristics of sound and their effect on people. Static pictorial storyboard [53] is often used to summarize the content of video files by presenting a collection of chronological keyframes [54] captured from the video. The contextual

visual summary design included in this dissertation summarizes and visualizes multimedia data through a combination of data cleaning/extraction and external links.

Network data can be visualized to present the connections and directions between different entities. The visualization itself is often a variation on the combination of matrix visualization and node-link diagrams [14]. It is also gaining more attention as people started exploring connection built through the Internet such as social media networks.

Many topics of interest can be presented and explored through some combination of the four groups of visualization techniques discussed above. For example, a collection of geotagged Twitter posts can be visualized by placing word clouds on top of heatmap hotspots to allow users to understand the major topics of conversation among different densely populated areas [55]. Most fields of interest that deal with a large amount of data and benefit from the use of visual analytics can be encoded with some combination of these four types of data. A typical combination used to tell data stories with more context utilizes numerical and textual data. Textual data are often overlaid on traditional numerical visualization techniques as (interactive) annotations [34, 35] to give reasons behind the numeric data's behavior. This approach is more effective when displaying a comparatively smaller amount of contextual data and can suffer from scalability issues.

This dissertation focuses on creating visual summaries for the four main data types in visual analytics, with the contextual data visual summary covering mainly textual data and a small portion of multimedia data. While the visual summary designs for each data type serve as an evaluation of the Three-Component Visual Summary designs, they also contribute as unique and function designs. The sections below review work relevant to the four applications.

2.4 Visualizing Time-Series Data

For the numerical data design, I applied the Three-Component Visual Summary to the most frequently used graphic design: the time-series plot [56]. In this section, I compare and contrast several time-series visualization techniques relevant to my design.

Familiar to most casual users, the line graph designed by William Playfair [20] is one of the most common statistical graphics [1] for time-series data. It displays the raw data in a simple and straightforward manner most can understand, meaning no additional training is required. It is measurable and easily comparable when the number of lines is small and the range of their values are close. However, as the number of lines and the range of their values increase, the graph becomes more complex and precise tasks become difficult as users start to experience cognitive overload [57].

Stack zooming [11] allows users to examine and compare focus points while retaining the overview context and provides visual clues to connect the two. My work also provides the ability to examine the details while keeping the overview in context, but as a visual display rather than an interaction function.

There exist more systems [58–62] that communicate large scale time-series data effectively through interactive exploration. My work, though not as effective as the systems, supports the exploration of time-series data when the interactive functions are not available.

Tree maps [22] are often used to display financial data [63], which is the primary time-series data examined in this use case. While it is powerful for displaying the hierarchical structure and trend of both the combined group and the individual commodities, it is not capable of displaying detailed changes over time which my work aims to also summarize.

The simple design of the band graph enables it to be a powerful tool in describing the overview of a dataset its audiences have no prior knowledge to [64]. The band graph is intuitive, with most of its components sharing the same appearance and functions with a traditional line graph. However, the band graph does not encode any information on the individual data sources. This means users will not be able to know the number of lines or the distribution of the lines, nor to identify any outliers. I learned from its design and utilize boundaries and a central value to communicate the overview of the data, while supporting the audience with additional analytical details of the dataset. Similar to the band graph, Fua et al. [65] introduced multi-dimensional graduated bands that encode the extent and the mean of polyline clusters in hierarchical parallel coordinates. This visualization technique

also allows, to a certain degree, comparison between different clusters. However, their work focuses on multivariate datasets, which are beyond the scope of this paper.

The stream graph [37] utilizes the ThemeRiver [66] layout to visualize the overall theme and its changes over time while preserving limited measurability on the individual lines. However, measuring the value of a data source now requires reading the height of its stream, rather than the y-value of the graph, which is slightly more difficult and confusing compared to a traditional line graph. Depending on the number of data sources, following the changes of a single data source can also be difficult as it does not have a stable baseline. Comparison between different data sources can also be difficult because of the increased difficulty in reading its value, and for having to compare between streams that may not be aligned closely. Though sharing a similar appearance, my work plots the lines using their true y-axis values, providing easier measurement and comparison.

The horizon graph [67] utilizes two-tone pseudo coloring [68] and separate charts for each time-series data to provide efficient comparison across a larger visual span [69] while preserving the movement of the individual commodities. This visual technique is capable of presenting a large amount of data and remaining readable. With each of the original lines normalized, the behavior of each line is clear and not affected by the overall scale, allowing users to compare the different data sources closely. By comparing all the individual data sources, which are now encoded by color, users can easily retrieve the overall trend, the overall correlation, and identify the outliers. As Javed et al. [69] stated, techniques such as this that create separate charts for each time series data provide more efficient comparison across a larger visual span than graphs that shares the same space such as line graph and band graph. Cloudlines [70] shares a similar design strategy to the horizon graph, utilizing separate and normalized space-saving design with the additional lens magnification interactive function to support a closer examination of the details. However, neither graphs' visual design provides value measurement, which is important to applications such as analyzing stock market data. My work, on the other hand, provides a simplified comparison between individual commodities on the important factors while retaining enough measurability.

Many charts precisely communicate one aspect of data but leave out the context that casual experts need to identify potential biases. For example, while treemaps communicate price and trend effectively, a user cannot determine how the comparison between different stocks changes over time and whether the trend is likely to continue through treemaps alone. Commercial tools, such as Tableau, allow trained analysts to explore datasets effectively by providing multiple instances of such charts, but are not designed to communicate the knowledge gained throughout exploration to casual experts efficiently. For example, online trading platforms often utilize visualizations such as the line graph (moving average, advance/decline indexes, etc.) and candlestick chart (high, low, open and closing prices) to allow their users to examine stocks and market indicators closely, generally one at a time using separate views. The sector or market summary is primarily visualized using graphs (treemaps, candlestick graphs, etc.) where users cannot identify detailed information for individual stock under the group. This requires users to obtain and compare the information between different views during different steps of interactive exploration to generate the insights desired with context and explanation. The research work presented in this dissertation focuses on efficiently communicating that knowledge to the casual experts by highlighting the important analyses while preserving and linking the analyses to the context in one single visualization. In the example of stock data presented in Chapter 4, for instance, the compact three-component visualization enables decision-makers to gain quick insight on the long term trends/highs/lows, indicators for short-term investment (e.g., sector increasing, but one or two stocks at low over a six month period) and for long-term investment (e.g., multiple sector stocks reaching all time high but showing downward trend indicating time to divest).

Aside from treemaps which do not encode time and techniques that rely heavily on interactive functions for exploration, I will compare visualization techniques utilizing the linear time structure [71] alongside my summarized line graph in their capability to complete different tasks in Chapter 4.

2.5 Presenting Academic Impact

For the contextual data design, I applied the Three-Component Visual Summary to a collection of documents, figures, and videos to visualize the impact of an academic-based research center. The dataset consists of primarily textual data and a small portion of multimedia data, as described in Section 2.3. In this section, I walk through research work on academic impact – a less defined field of interest for casual-expert decision-makers that often deals with unstructured textual data collections – and explore how academic impact has been visualized.

Eugene Garfield first introduced the idea of impact factor [72] in 1955 from the study on Citation Indexes for Science. It eventually became a tool many use to evaluate scientific achievements. Scientists and published papers within certain fields of study were ranked and compared using the impact factor. The impact factor is, however, often misused [73], and is more useful in comparing researchers or research works rather than understanding the impact and its dynamics. It is also limited to achievements related to publications, and academic impact should consider more than just publications.

Borner and Scharnhorst [74] reviewed different science conceptualizations used for comparing existing datasets and models. These visualizations, however, mostly focus on the relations and the connections between different topics or specific works and are not fitting for visualizing the development of impact from a specific center.

CoE-Explorer [75] was created to help DHS program managers analyze and present the research works within DHS' network of Centers of Excellence, led by multiple universities, to external policy makers. The system visualizes details of individual centers, projects, and investigators. Its primary output, however, provides the general themes as an overview of the different works within the research organization, rather than a better understanding of the research works and their impact.

STAR Metrics [76] is a repository of data and tools that allows users to examine the funding and impact of federal investments across the United States. It allows users to drill down into the data by location, organization, topic, project, etc. The visualization

tools serve more as filters and the data remains to be displayed in a spreadsheet and text description format. The impact unfortunately only considers the final list of publications and patents.

Similarly, Madhavan et al. [77] created DIA2 to help casual experts, specifically program managers and academic staff at the National Science Foundation, explore research funding portfolios. The system also uses a combination of structured and unstructured NSF-related data and focuses on organizational structures, collaboration networks, fundings and awards received. Users can get a glimpse of the impact of a particular portfolio over the years through simple statistical graphs such as a bar graph on the number of awards received, the amount of funding received, or the number of collaborators per year. However, DIA2 focuses mainly on the final result and presents minimal contextual data in the storytelling of the temporal evolution. DIA2 also has less support for the analysis and comparison of different impact variables within a portfolio or between different portfolios.

2.6 Geospatial Visual Analytics

For the geospatial data design, I applied the Three-Component Visual Summary to visualize crime reports from Tippecanoe County, Indiana, to support decision making in the law enforcement field. Multiple geospatial visualization techniques were implemented into the design. This section first covers the background, the challenge, and the opportunities with geospatial visualization, then discusses the different geospatial visual analytics designs relevant to my application.

The history of geospatial visualization can likely be traced back to the early data maps such as the Yu Chi Thu from the eleventh century [56]. Markers and annotations were added to maps to help people better understand the physical characteristics and noteworthy knowledge of the environment. There were even three-dimensional geospatial visualization tools such as a tactical sand table that help military leaders strategically place or move their resources.

However, the more modern geospatial visual analytics that combine cartographic and statistical methods to create data maps started around the seventeenth century [56]. Some of the more significant early works include Edmond Halley's chart for trade wind and monsoons [78], John Snow's map on Cholera [79], and Charles Joseph Minard's maps on French wines and Napoleon's army in Russia [80]. The majority of these visualization techniques encode a variable directly on top of a map to help the audiences identify location-related patterns. Halley's chart overlays magnitudes and directions above the area that represents the sea, Snow's map aligns bar charts on the count of death along the streets, and Minard's map added lines with various width or pie charts on top of different locations on the map. Instead of showing accurate geo-coordinates on a map, some geospatial visualizations also distort the distance or the final location to allow a more understandable communication while maintaining a basic understanding of the relative location [30, 81]. Some geospatial visualizations, on the other hand, utilize a three-dimensional space and the additional z-axis to encode the additional variable [82, 83]. However, these three-dimensional visualizations can often be difficult to use and require more effort to change the viewing angles for use.

Nowadays, geospatial visual analytics are often used to provide insights into clusters and movements. Rather than simply encoding the location of one contextual variable onto a map, geospatial visual analytics today often encodes data of other types such as time [83] and text [41–43]. Such insights can help first responder analysts better analyze and predict the locations of events such as crimes [44], boating accidents [45], and so on. Different spatial correlation or clustering calculation methods such as Moran's I [84], Geary's C [85], Getis-Ord General G or Getis-Ord G_i* [86] were also developed to measure the different spatial characteristics and are supported by tools like GeoDa [87] or ArcMap [1]. However, these tools often require many input variables and can be overwhelming or confusing for those who are not familiar with the methods.

On the other hand, VisMaster urged designers of geospatial visualizations to support a broader community of potential audiences that may not be trained in data analysis with

[1] http://desktop.arcgis.com/en/arcmap/

”lightweight, easily deployable and usable software that allows flexible customization and combination of tools” [88]. This work attempts to focus on that challenge and opportunity.

Djavaherpour et al. [89] presented a physical 3D model of earth to support geospatial analysis. Users are given a selection of attachable styling and analysis layers to explore different datasets and characteristics of geographical regions. This approach provides flexibility in an environment that is intuitive and requires no technical expertise to operate. Further, this approach allows detailed analysis while keeping the overview in sight. While it can be challenging to explore either a scale or an analysis that is not printed, generating and printing new layers to add to the model is straightforward. However, the physical model limits users in transporting and sharing the tool. It also does not utilize the computational power of the machine once the model is printed. This work utilizes a similar approach in providing pre-generated layers of analysis results while retaining certain aspects of the overview but allows users to update the attribute of the stacking layers to achieve a balanced visual outcome. The ability to easily capture and share the visual summary is also important to this dissertation.

Similar to the physical 3D model, Wagner Filho et al. [90] brought the navigation of geospatial trajectory data into a 3D environment to address the lack of depth cues and reduce the learning curve of examining 3D data in a 2D space. Immersing users in a 3D space opens more design space to encode variables and attributes. However, the set up requires specific hardware that is not yet commonly owned. This dissertation shares the same goal as both Wagner Filho and Djavaherpour in reducing the learning curve required to explore the geospatial data, but focuses on using mediums that are accessible to most.

Van Ho et al. [91] introduced a framework to shorten the time and effort in developing customized web applications for geospatial visual analytics. The framework allows users to interactively explore and analyze geospatial data and publish vislets to communicate the finding. However, the published storytelling is heavily author-driven and can lack the necessary components for context or additional analysis. This dissertation shares the goal of minimizing the effort to communicate the data using a customized visualization, but tries to balance more evenly between author-driven and reader-driven stories.

Godwin et al. [92] introduced TypoTweet Map, which replaces map features with texts of different colors and sizes to relate the collection of texts from social media to spatial locations. This design brings the opportunity to provide more context in the base map, which can be beneficial to the data envelope component in my geospatial design. However, the amount of text displayed could be overwhelming to a user when a substantial amount of geographical region is covered. This can make the design inaccessible to a casual expert.

Godwin et al. [93] also used path KDE (Kernel Density Estimation) to visualize geospatial data in urban spaces. By drawing paths with different thickness and saturation along streets, the spatial distribution of data can be visualized in a manner similar to heatmaps without compromising the clarity of the street map. Maintaining the clarity of the base map while encoding geospatial attributes can be useful to the layering approach in my proposed geospatial design. However, this technique relies on the presence of streets and could be more limited for rural areas.

Slingsby et al. [94] worked closely with animal movement ecologists and introduced a set of requirements for designing interactive visualization systems with the targeted audience of domain experts. The design requirements include providing access to the original data, allowing exploration at different scales, requiring as few user-interface interactions as possible, and including contextual geographical data (such as landscape) to help interpret spatial data. This dissertation shares a similar strategy in designing visualizations for casual experts, specifically in the provision of raw statistics and additional spatial context. However, this dissertation also focuses on creating a presentable coherent visual summary to address time limitation.

MacEachren et al. [95] introduced a map-based interactive application to support crisis management with Twitter data utilizing a multi-view interface which includes a map view, a tweet list, a time-plot/control, a query window, and task list that was still under implementation at the time the paper was published. The map view provides an overview of the dataset using a gridded density surface and the individual data entries with selectable points simultaneously. The specific content of the Tweets is placed on the side with the selected post highlighted. Users can filter the data displayed through a time bar (that also serves as

a frequency plot) or text-based queries. However, the system only provides basic query-based analysis. It focuses more on identifying interesting data entries while my geospatial design focuses more on analyzing and comparing the spatial characteristics of data subsets.

The rest of the section discusses specific geospatial visualization techniques that can visualize large amounts of geo-tagged data over a map view, making them appropriate for my application.

Scatter plot [36] that uses the geo-coordinates as x- and y-axis can faithfully represent the entire dataset with the option to use color, radius, or shapes to encode additional variables. However, labeling or displaying overlapping data points can be challenging in the scalability aspect.

Density function visualization [28], or sometimes referred to as a heat map, can summarize the density of the data points on top of geospatial regions. This is especially suitable for use cases focusing on whether or not a location has a data point present, contrasting use cases that focus on the content of the geo-tagged data. Density function visualization also handles the overlapping points and provide users with a more accurate visual representation of the data distribution. Contour maps can also serve as an alternative visual representation option of the density function visualization. While contour maps can require more effort to precisely understand the data, it occupies less space and allows more additional information to be displayed on the map.

Choropleth maps [96] focus more on the patterns and the changes of data content throughout the different geographical boundaries rather than the actual data distribution. The sub-regions are usually colored based on the content and the count of the data within. A common practice is to select the hues based on the content and the saturation based on the normalized count. While this approach cannot reflect the distributions of data within the sub-regions, it does shorten the time to digest the visual representation, and users can often adjust the scale of the sub-region to examine finer details. A quadtree-based approach introduced by Thom et al. [97] with a user-defined threshold can define sub-regions that present the distribution of the data better but may interfere with geography-based boundaries, which could encode additional context important to users. Instead of coloring the

sub-regions, some visualization techniques overlay different simple statistical visualization techniques on the different sub-regions such as pie charts or bar charts for data that encode more variables. Such techniques, however, can block the map tile information and may be difficult to be co-presented with other visual components and are not considered in this design.

The methods described above, while having different strengths and limitations, share the same geospatial space in a more compatible manner, and thus can theoretically be combined in the same visualization. Finding the balance and the designs that accommodate each other will be the key to effective information transfer and insight generation.

2.7 Network Visual Analytics

For the network data design, I applied the Three-Component Visual Summary to visualize data flow between different entities and communities to support first responders in the understanding of information source and influence. The design focused on directed, non-weighted, (ground-truth) community datasets such as a Twitter following network. This section discusses network visualization approaches relevant to my application.

While modern social network analysis can be traced back to the late 1920s [98], the advancement of the internet widely expanded the possibilities of social networks [99] and sparked more interest in the insights that can be discovered through social networks analyses [100]. Since every network can be represented as a graph [101], visualizing and analyzing networks using graph theory became a common practice [102, 103]. A network, therefore, is often visualized as a node-link diagram and occasionally as an adjacency matrix [104, 105] or some combination of the two [14].

Additional graph visualization layouts that can represent a complex graph include Hive Plots [106], ARC Diagrams [107], Sankey Diagrams [108], Chord Diagrams [109], and Pivot Graphs [110]. Out of these visualization techniques, Hive Plots, Chord Diagrams, and Pivot Graphs require additional domain-specific data to visualize [103]. Arc Diagrams visualize the data over a one-dimensional layout, which can suffer in conveying the overall

structure of the graph and using space effectively. Sankey Diagrams focus on the flow quantity and mainly support energy, material, or cost transfer visualizations.

While the majority of the designs utilize node-link diagrams for the more straightforward visual representation and easy-to-follow connections, the scalability decreases significantly as the connectivity/number of edges increases and creates visual cluttering [111]. Interactive functionalities and layout designs were introduced to address the challenge of scalability [112–114]. These solutions focus on giving readers access to the information of interest, but do not solve the remaining visual clutter, and may require the audience to know the nodes or edges of interest before modifying the layout. These solutions also require the position of the nodes not to encode data that cannot be altered, such as relative geospatial coordinates or schematic overlays. While this work focuses on network data with flexible node positions and is not constrained by such limitations, it aims to give its audiences the flexibility to alter the layout as desired to highlight the audiences' interest without depending on extensive interactive exploration to extract and understand the remaining data, which is difficult to satisfy concurrently by the solutions mentioned above.

An adjacency matrix, on the other hand, is an effective way to visualize a dense network [115] as it removes the potential overlapping of the nodes and the edges. However, path-related tasks, which could be crucial for network analysis, become less intuitive compared to the node-link visualization [116]. Adjacency matrices can also be confusing when encoding directional connections to those unfamiliar with the visual representation, which can be a problem as this work focuses on directed network data, meaning it is important to be able to present the direction of the data flow to our target audiences who may not be knowledgeable with the visualization techniques. Finally, the overall global structure of the dataset can be hard to identify with the adjacency matrix representation. On the other hand, a node-link diagram has the potential to encode more variables with its flexible node positions and potentially varied style, length, and width of the edges than the matrix representation. This work tries to combine the strength of both designs and overlay additional visual elements to include additional information.

Side-by-side use of the node-link diagram and adjacency matrix can allow users to utilize both designs effectively switching between the two based on the task [117]. However, this strategy uses additional space to display the same information twice, and often require users to create the mental connection of the two displays through synchronized changes when one design is being updated, which means a heavy reliance on interactive functionalities which this work tries to reduce.

Clustering entities of similar properties and display them in close proximity visually can also help users navigate through a node-link diagram more easily by reducing the number of crossing edges [118] which also improves the aesthetics of the drawing [119]. Different algorithms were created to automatically identify clusters of nodes or edges that can be visualized with less crossing edges [120–122]. However, with the goal to better preserve context, I chose to work with ground-truth communities that are pre-defined based on real-world context rather than the connections of the entities. This work reduces the number of crossing edges by replacing the node-link visualization within each community with the adjacency matrices instead and only preserving the edges that connect different communities. Rather than aggregating the cross-community edges with the same source and destination communities to create a cleaner appearance similar to Auber et al.'s multiscale visualization of small world networks [123], my network design retains the underlying links to allow users to trace the paths more fluently and preserve node-level context.

Nodetrix [14] is a hybrid visualization for large social networks designed by Henry et al. How the design uses a combination of the node-link diagram and adjacency matrix to present simultaneously the global structure and the community detail echos with the direction of the Three-Component Visual Summary design. However, Nodetrix targets weighted, non-directional, community network data and does not support direct extractions of more complicated network analysis results in its visual design. I decided to build my network application on top of the NodeTrix design with the addition of overlaying analytical highlights, contextual supports, and adjustments for the different data characteristics. Similarly to Nodetrix, this work uses force-based layout [124] to calculate the initial positions of the matrices but allows users to adjust the position manually.

The rest of the section discusses and compares relevant visual analytics work for network data.

Frishman et al. [125] presented an algorithm for visualizing dynamic clustered graphs which maintains and separates the communities within the network dataset. This design provides an overall structure similar to NodeTrix by highlighting each community with colors and boundaries but retains the node-link diagram within each community. While path-based analysis within the community may be more straightforward, the intersecting edges can create additional visual clutter. While both approaches are capable of incorporating the overall structure and the detailed connections in the visualization, my network design builds on the NodeTrix layout for a cleaner visualization and utilizes machine computation to reduce the effort needed for path-related tasks.

Chen et al. [126] introduced a web-based system to support their structure-based suggestive exploration approach. The system utilizes multiple views to present the overview, detailed network structures, and the exploration history simultaneously. Users can see the selected area for the detailed node-link diagram in the heatmap that represents the overview to maintain the perspective. The exploration history attempts to capture the interactive exploration in a static presentation and also serves a selection tool. The majority of the queries can be accomplished through click and drag, reducing the learning curve of the system. However, the exploration history presents each historical moment with a representative high-level structure and does not capture the "why" between the moments. The system also focuses on finding structures and has little support for path analysis. My network design also keeps users informed of both the overall structure and the detailed connections and focuses more on analyzing the data flow.

Major et al. [127] introduced Graphicle to address the analysis opportunities and insights that are potentially missed when focusing primarily on either the underlying network structure or the individual data units. This dissertation also focuses on ensuring data entries relevant to decisions are not unintentionally left out. However, Graphicle is interaction heavy, and the packed unit visualization method can be challenging to understand and explore without the proper training. In the study presented, participants were given

20 minutes of training and 30 minutes of free exploration before being tasked. This dissertation wishes to present tools that are less dependent on interactive explorations for the decision-makers.

Wong et al. [128] presented a graph analytics model that explores a network starting at the middle-ground information that is often overlooked by the top-down and the bottom-up approaches. This approach separates a network into hierarchical node-link diagram layers, each with a different level of detail displayed. By starting in a middle layer, users are given the overall structure of the network and enough details to identify areas of interest. Users are then given the option to expand or collapse parts of the network to examine entries of interest while retaining the overall structure of the dataset for context. This approach allows users to examine a specific detail of interest while keeping the structure of the dataset in perspective. Similarly, my network design also provides an overview and the details in one visual presentation. However, my network design displays the connections of the "collapsed" nodes in the adjacency matrix representation and incorporates a set of analyses into the visual design. In comparison, Wong's approach focuses on examining and traversing large datasets. Wong et al. [129] further explored web-scale graphs through a peek-and-filter strategy with data preprocessed using GEM. To be able to output manageable static visual summaries, my network design focuses on datasets at a smaller scale. However, my design also pre-compute various analyses to allow a smoother experience under the processing power of web browsers.

Perer et al. [130] developed MatrixFlow to visualize the temporal flow of network data. MatrixFlow visualizes temporal evolutions of network data through a series of aligned adjacent matrices. This approach visualizes information traditionally gained through interactions or animations over multiple visual components to present the story behind the data in a coherent visualization. However, this design focuses mainly on aggregated change over time and offers little support for path-based analysis. This dissertation also utilizes multiple visual components to present information traditionally gained in different stages of interactive explorations. However, the proposed network design utilizes the Nodetrix layout, which is capable of incorporating more layers of visual analysis representations.

Srinivasan et al. [131] introduced Orko, which allows both natural language input and touch-based direct manipulation input. This approach could potentially reduce the technical expertise required for casual experts to operate visual analytics systems. Orko also fades out (rather than removing) the filtered data entries and provides a summary container of the selected nodes. Similarly, my network design provides community-level summaries and moves the details to the background. However, Orko focuses on interactive exploration, while this dissertation focuses on communicating knowledge through a coherent visual summary.

3. THREE-COMPONENT VISUAL SUMMARY

A version of this chapter has been previously published in Computer Graphics Forum.
Citation: C. Yau, M. Karimzadeh, C. Surakitbanharn, N. Elmqvist, and D. S. Ebert, "Bridging the Data Analysis Communication Gap Utilizing a Three-Component Summarized Line Graph," *Computer Graphics Forum*, vol. 38, no. 3, pp. 375—386, 2019.
Doi: 10.1111/cgf.13696

The large volume of data we are able to collect and generate with modern technology can be both a blessing and a burden. On the one hand, the more data we have, the more likely it is to embed patterns valuable for new insights and predictions. On the other hand, it requires more effort for the analysts to obtain a comprehensive understanding of the dataset. This is a greater challenge for the targeted audience of this research work – domain experts who have little training in information retrieval or visualizations – even with the aid of tools like visual analytics. In this chapter, I propose Three-Component Visual Summary, a visual summarization design that infuses that exploration process into the visual design to support casual experts in making data-driven decisions while minimizing the dependency in interactive explorations.

3.1 Design Requirements

The reliance on interaction, animation, and multiple/larger displays during data exploration has been increasing in visual analytics practice. Because of the complexity of the data collected, understanding and analyzing data typically involves several interactive steps, such as overview, zoom, and filter [2]. This kind of visual exploration process can be an effective way for analysts to gain first-hand understanding and insight into the data. However, the tools and techniques designed for use by analysts do not preserve informa-

tion pertaining to how the analysts reached their conclusions. An individual who did not participate in the data analysis process only views the final results without the explanation of the process. Consider a stock market analyst trying to make a recommendation for the "air freight/delivery service" subsector of the Nasdaq transportation industry. In order to satisfy the goal of picking the best stock to invest in, the analyst may progressively filter out competing stocks based on the analyst's acquisition guidelines until the analyst has isolated two specific stocks that the analyst sends to the manager to approve. Without the context of the full analysis session, including the possible hypotheses tested and discarded, the manager has limited understanding of the overall findings, no way to detect potential information bias, and no recourse to check the work. This gap in the understanding of the data could prevent the manager from making optimal decisions.

Therefore, I draw on the concept of communication-minded visualization [5] to bridge and minimize this gap between data analysts and their final audience (e.g., managers, decision-makers, the general public) by incorporating the previously missing but important contextual knowledge into the design of novel visual summaries. To achieve this, I first have to understand how data analysts transfer the obtained knowledge to the decision-makers. To gain this insight, I surveyed decision-makers from the first responder community (e.g., public safety, police, rescue) on the way data is prepared and presented to them by data analysts. Through the understanding of these current practices and the needs of this audience, I compiled a list of requirements for communicating data that focuses on increasing efficiency, improving understanding, and reducing the impact of potential information bias:

R1 **Comprehensibility:** Decision-makers must be able to quickly acquire a basic understanding of the overall behavior.

R2 **Accuracy:** Decision-makers must be able to efficiently and accurately identify insights that are important to a decision.

R3 **Fidelity:** Decision-makers must be given the ability to explore and understand the original dataset and the reasoning leading to the analytical highlights.

R4 Precision: Decision-makers must be able to obtain actual data values.

R5 Comparison: Decision-makers must be able to compare significant aspects of the subsets of data.

Based on these requirements, I designed the *Three-Component Visual Summary* which visually summarizes datasets using : (1) representative data, (2) analytical highlights/comparisons, and (3) a data envelope. The representative data provides the audience with a quick overview of the entire dataset, the analytical highlights/comparisons allow the audience to generate insights of interest with ease, and the data envelope summarizes the remaining aggregated data to enable simple exploration of the raw statistics. The Three-Component Visual Summary is designed to be a method that can be applied to different data types and incorporate different visualization techniques where the designer identifies the three components, finds the appropriate visual representations, and creates the visual summary.

3.2 Survey: Communicating Insights to Stakeholders

This work was inspired by regular interactions with decision-makers in multiple projects spanning an extended period of time. Based on the feedback collected during these interactions, I designed a survey to understand the problems inherent in communicating insights from analysts to stakeholders. Below is a summary of the survey findings. The complete survey questions and responses can be found in Appendix A.

F1 Data analysts are given limited presentation time

F2 Decision-makers are given limited exposure to data

F3 Data analysts and decision-makers understand the data differently

F4 Information bias can exist

F5 The presentation has the power to influence the decision

3.2.1 Survey Method

To better understand how data is communicated between data analysts and decision-makers, I surveyed six decision-makers from the first responder community. The participants, who represent my primary target user community, are decision-makers at different levels in police and public safety departments.

The survey was conducted in March 2018 over Google Forms with the results collected anonymously. The survey contains a total of eight questions: six multiple-choice questions on the different characteristics and expectations of the presentation experiences, and two short response fields for explanations. While I reached out to eight decision-makers, only six results were submitted.

Of course, with the limited sample from one specific group, the result may not represent all practice; I have, however, found similar needs in a survey of financial analysts [132]: providing context, supporting analyses, allowing comparisons of the details, etc. From my surveys, I concluded a few key points in current practices that identify the limitations of this communication process between data analysts and decision-makers and derived the design requirements in Section 3.1. These results support the feedback I have received from many decision-makers in the past.

3.2.2 Survey Findings

In this subsection, I expand on the survey findings and elaborate on the implication for the design requirements.

First, common among all respondents was a limitation that *data analysts often only have a limited amount of time to present their findings* (**F1**). While decision-makers at different levels and in specific fields have different practices, analyst presentations are often limited to five minutes or less. The short amount of time means the data analysts can only communicate a limited amount of data, and decision-makers cannot afford to waste time on results that are not important. This ties into the requirements of my visual design **R1** (comprehensibility) and **R2** (accuracy), where the decision-makers must be able to quickly

and efficiently create an understanding of the dataset and its highlights. This time limitation also constrains the presentation to focus on the dataset and the insights instead of a full exploration of the exploration process.

I also found that *decision-makers only have limited exposure to the data* (**F2**). It is clear that with the limited amount of time analysts are not likely to be able to walk the decision-maker through the entire dataset. From the survey, none of the data analysts always include raw statistics in their presentation and one-third of the data analysts almost never include raw statistics in their presentation. On the other hand, all of the decision-makers acknowledge the impact that seeing raw statistics has on their decision making and like to see the raw statistics at one time or another. Two of the six decision-makers actually wish to see raw statistics at all times. It is, of course, unpractical to present all the data with the limited amount of time, but as one of the decision-makers stated: "[The data analysts] have [the data] in volume. I need it in highlights with the ability to ask for more." The decision-makers should be given the ability to better understand the raw statistics, which ties into the requirement **R3** (fidelity).

Data analysts and the decision-makers understand data differently (**F3**). One of the decision-makers stated that "[data analysts] tend to focus on the manner in which data is captured whereas [decision-makers] tend to focus on the story the data is telling." While that may not hold true for all data analysts, it is not surprising that important details may be lost during the filtering process because the data analysts have a different focus in mind while exploring and preparing the data for presentation. For example, when presenting a dataset with just the average and standard deviation, anomalous spikes in a specific data source that lasted only a short amount of time can easily be overlooked and not presented. This ties into the requirements **R2** (accuracy) and **R3** (fidelity) in which decision-makers must be able to identify insights that are important to their decision and explore the data to a certain extent instead of letting the data analysts have complete control over deciding what is important in the presentation.

Additionally, *there can exist information bias* (**F4**). Five of the six decision-makers have experienced situations where presented information appeared to be biased toward a

decision. Note that sometimes data analysts also present the preferred courses of action which ideally align with the data when conveyed appropriately, even if that counts toward providing biased information. However, as stated in the work of Ajzen et al. [133], personal relevance could affect what is viewed as the preferred outcome even under the same work field. As a result, information bias can happen when data are interpreted differently between data analysts and decision-makers, leading to an error in the conclusion. This leads us back to the requirement **R1** (comprehensibility), **R2** (accuracy), and **R3** (fidelity) where the decision-makers must be able to obtain an understanding of the dataset that is enough to evaluate the presented options objectively to minimize the impact of information bias.

Also, *data influences real-world decision making* (**F5**). All of the decision-makers from the survey acknowledged that the data at least sometimes affect the outcome of their decision. While this may seem obvious, it is important for the data to be measurable and comparable to allow the decision-makers to link the data to additional real-world variables for decision making. This ties into the requirements **R4** (precision) and **R5** (comparison) where even in a summarized visualization, decision-makers must be able to measure and compare significant factors relevant to their decisions.

Finally, *the presentation can be limited by its medium*. With limited time (**F1**) and different settings, it has been observed that analysts are sometimes limited to presenting the processed result (**F2**) using static images. Presenting such static charts means the final display should be self-contained, i.e., it should incorporate the relevant analysis results without the need for interaction or animation. By summarizing the dataset visually and including noteworthy insights, the cluttering on screen and the dependency on interaction or animation can be reduced, allowing the decision-makers to retrieve the same information more efficiently even with just static images.

A summary of how the survey findings are addressed by the design requirements can be found in Fig. 3.1.

3.3 Component Design

To summarize the data for efficient knowledge retrieval without losing important details and address the design requirements, the summarization design is driven by these characteristics derived from the requirements from the survey:

First of all, the visualization must present a summary of the data to satisfy **R1** (comprehensibility). The audience being presented with the data needs to be able to quickly understand the basics of the dataset, and therefore representative values will need to be clear to users at the moment the data is being presented. Since summarizing and focusing on the main takeaway can result in losing perspective on parts of the dataset, it is important for users to be able to obtain a basic understanding of the scope and the distribution of the actual data to satisfy **R3** (fidelity) and **R5** (comparison) even with the visualization focusing primarily on the summarized components. Additionally, to minimize the impact of information bias and satisfy **R2** (accuracy), the visualization should not only enable quick extraction of important analysis results, but also allow its audience to easily understand how these results are generated from the data. Finally, to satisfy **R4** (precision), the above need to be measurable.

To include all of these characteristics, I separate the data into three components in the final visual design, each with a different focus and priority in encoding the data, and combine them to provide an improved and balanced visual summary presentation. The three components are: *Representative Data*, *Analytical Highlights*, and *Data Envelope*

The Representative Data provides the audience with a simple but precise description of the dataset (**R1**). It should be clear and easy for the casual users to understand without additional training and should be able to be communicated quickly. Visually, the representative data should be the most prominent element in the visual summary.

Analytical Highlights are added to the visual summary as the second visual component to reduce the time needed to gain useful insights from the dataset, to ensure it is clear how the insights are extracted, and to minimize the loss of important discoveries during the exploration (**R2**). The highlights should provide insights that are not straightforward for

the viewer to identify directly from seeing the raw data but are important to the outcome of the data exploration. This component should be designed to address the specific insights of interest to the decision-makers and allow them to compare different aspects in the dataset for decision making (**R5**). In the visual design, the analytical highlights should be easy to identify and provide the connections to the raw data to support the analysis results. Visually, the analytical highlights should not outshine the representative data, but should remain easily recognizable.

The Data Envelope summarizes the remaining aggregated data to put the first two components into context (**R3**) and therefore aids the decision-makers in understanding and evaluating the options and conclusions provided by data analysts despite potential information bias. It should provide simple yet specific (**R4**) details (e.g., boundary values) of the raw data that are not included in the representative data, and possibly allow basic comparison between different data points (**R5**). Visually, the data envelope should be less prominent compared to the representative data and the analytical highlights, so it provides context but does not distract. It should also be presented in a simplified manner to reduce the complexity of the visualization.

The three components are then combined into a display to create a visual summary. The three components should be presented in a way that encourages the audience to first examine the representative data, the analytical highlights/comparisons, then finally the data envelope. By displaying all three components simultaneously, the visual summary can directly present knowledge that is traditionally retrieved at the different levels of exploration to address casual experts' limited time, and reduce dependency in interactive exploration to address casual experts' lack of training in visual analytics. By having the overview and the low-level details present when examining the analytical highlights/comparisons, this design can preserve the context and therefore reduce the impact of potential bias from the data analysts. The three components can be presented simultaneously by directly overlaying them on top of each other or by using multiple views, which I will demonstrate in the designs for the different use cases.

To accommodate the casual experts' lack of training in data analysis, this design should utilize familiar visualizations and statistics to create the visual summaries. A summary of how the design requirements are incorporated into the design of the three components can be found in Fig. 3.1.

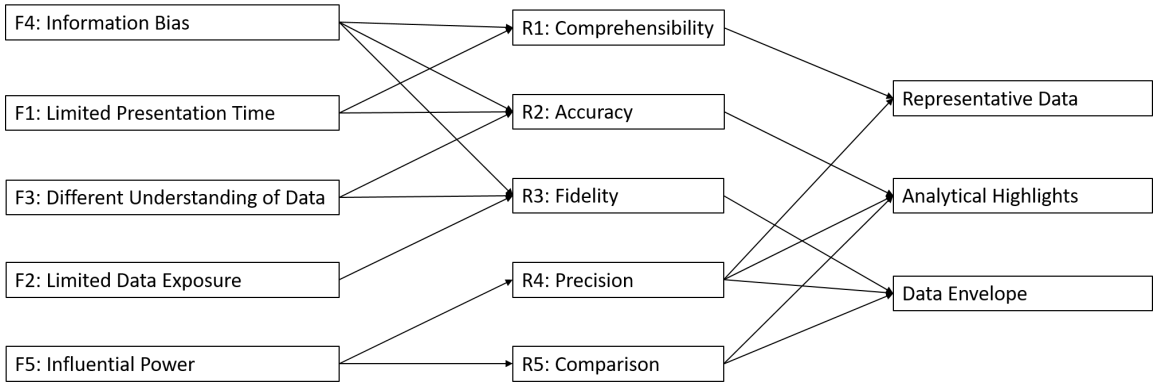


Fig. 3.1. A summary of how the three components satisfy the design requirements and how the design requirements address the survey findings.

3.4 Constrained Interaction

While the Three-Component Visual Summary design aims to incorporate scenarios where only static images are being used for presentation, the proposed three-component layout can be applied to an interactive environment to support data analysis and the generation of the static images for presentations. To reduce the hindrance from lack of training in visual analytics tools and to provide a better storytelling, the Three-Component Visual Summary designs support users with constrained interactions [7]. Considering the designs can be used to present the data to others, I limit the interactive functionalities to the level that is intuitive on a touchscreen tablet, such as using click or hover to highlight, rearrange, or display annotations in a tooltip. The visual summaries should utilize simple representations and common signifiers familiar to the audience in the display of the interactive functionalities such as the save icon or the "click-able" mouse cursor icon.

3.5 Strengths and Challenges

While traditional summarization techniques focus on the aggregation of data and present all of the data entries equally, the Three-Component Visual Summary design identifies and highlights the data entries relevant to the decisions to be made. The Three-Component Visual Summary design constructs a visual summary using three different components to incorporate insights traditionally gained from the different levels of interactive exploration. This design provides the audiences with knowledge similar to that retrieved following the Shneiderman mantra [2] in a shorter amount of time. The flexibility to choose the variables for the three components also allows the visual summary to be a more customizable and focused experience, which works toward reducing the communication gap.

However, the effectiveness of a three-component design in helping a decision-maker generating insights relevant to the decisions now depends on the designer's ability to select the appropriate variables and visual encodings to construct the visual summary. While there have been many studies on the strengths and potential harms of different visualization techniques, layering multiple visual components will introduce new challenges to be addressed. The fact that the same visualization techniques can be used differently in different Three-Component Visual Summary designs for different use cases or datasets can also complicate learning multiple Three-Component Visual Summary designs.

My attempts in applying this visual summary design, which will be explored more in the following chapters, suggest that this approach can enable casual experts to extract a more accurate and extensive understanding of the dataset in a shorter amount of time compared to existing visualization techniques. I believe this is a successful first step toward verifying the potential of the Three-Component Visual Summary design in bridging the data analysis communication gap.

While the initial attempts focus on the specific data types and use cases, I believe the Three-Component Visual Summary design can be applied to more data types and applications as long as the designer can identify components that satisfy the characteristics de-

scribed in Section 3.2.1. The scope of this thesis work focuses on the four major data types in visual analytics, which will be described in more detail in the following chapters.

4. SUMMARIZING NUMERICAL DATA

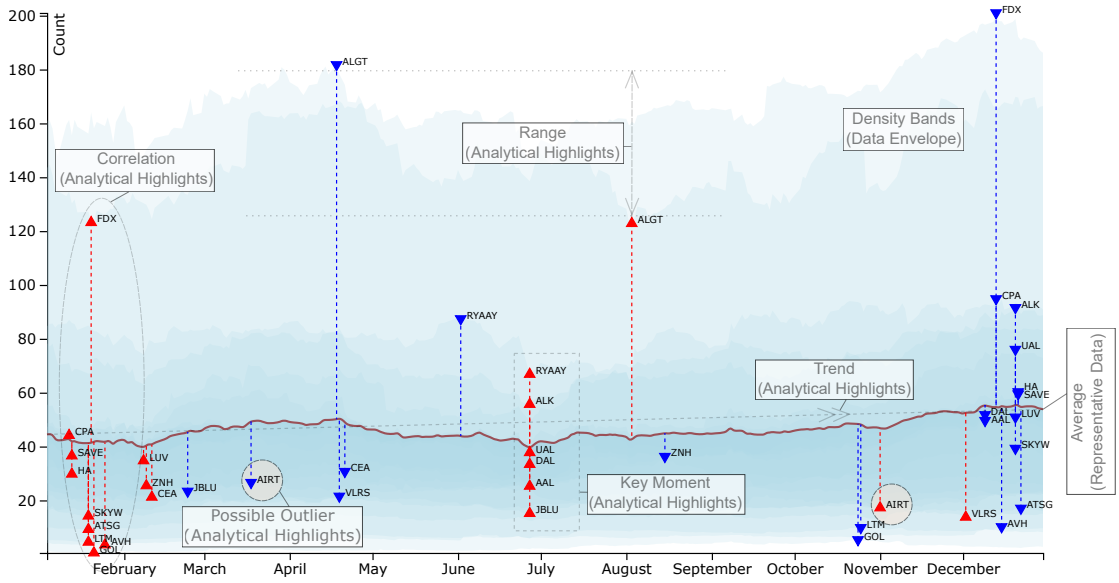


Fig. 4.1. Example of the three-component summarized line graph showing Nasdaq stock prices of the transportation industry in the air freight/delivery service subsector during the year 2016 (a total of 21 stocks over a full year). The tan line is the *representative data*: an average curve providing the mean value for the entire summarized dataset. Along the time axis, *analytical highlights* are shown as ranges, trends, correlations, outliers, and key moments called out using dotted lines and triangles; red triangles represent the absolute minimums of each line, and blue triangles the absolute maximums. Finally, the light blue bands in the background provide the *data envelope* that give the data distribution over the entire time axis.

A version of this chapter has been previously published in Computer Graphics Forum. Citation: C. Yau, M. Karimzadeh, C. Surakitbanharn, N. Elmqvist, and D. S. Ebert, "Bridging the Data Analysis Communication Gap Utilizing a Three-Component Summarized Line Graph," *Computer Graphics Forum*, vol. 38, no. 3, pp. 375—386, 2019.

Doi: 10.1111/cgf.13696

I created a summarized line graph following the guidelines proposed in Chapter 3 as the first step to evaluate the Three-Component Visual Summary design. The summarized line graph, as shown in Fig. 4.1, features the mean curve as the representative data; range, trend, correlation, outliers and key moments as analytical highlights; and extent, density, and traces as the data envelope. I selected a line graph because it is one of the most commonly used and readily understood visualization techniques for presenting simple time-series data [1]. This simplicity allows this technique to be used to support decision making in multiple domains. For example, in the finance field, investors can compare the performance of stock prices between different companies and evaluate the time and the stock to invest in or to sell out. With the same graphical design, decision-makers in the field of public safety can examine counts and trends of different incidents over different regions to reevaluate the budget and resources allocated to the various teams. Since there are many visualization techniques that focus on the efficient analysis of time-series data [134], I can use these applications to evaluate the effectiveness of the summarized line graph. This proposed visualization technique targets casual experts, e.g., decision-makers with strong domain knowledge but limited time or training for advanced visualizations. The proposed technique is able to effectively and efficiently communicate multiple quantitative time-series data and their correlations. I first explain and discuss the design choices for the three components in the summarized line graph and then the process of constructing the final display using the three components. I compared the performance of the summarized line graph against traditional line graphs, band graphs, stream graphs, and horizon graphs in a user study on both their complexity and ability to meet the requirements listed above. The study described below measured performance in terms of both accuracy and completion time for four representative tasks drawn from the requirements: identifying the original data, the overall trend, the outliers, and the key moments. While the results indicate summarized line graphs do not outperform other techniques in every task, it achieves the best overall result when all four tasks are considered.

4.1 Design

In this section I describe the design of my summarized line graph and how this design technique ties into the requirements that motivated the Three-Component Visual Summary design.

4.1.1 Representative Data

The summarized line graph plots the average of the data as the representative data. The average (mean), being one of the most used descriptive statistics, is capable of providing measurable values (**R4**) and change over time that represent the central tendency of the entire dataset (**R1**) and is easily understood by casual users. Summarizing the entire dataset with one line on the graph creates the initial focus in the visualization, which provides a simple but effective visualization.

I choose to use mean over median to focus on values rather than order. While mean is susceptible to the influence of outliers and can be misleading when extreme values exist in the dataset, the data envelope is designed to counteract this problem. Additionally, the sudden jumps in median values per time-step create more abrupt changes in the visual displays and hinder trend analysis.

4.1.2 Analytical Highlights

The analytical highlights component is a component that can be customized to fit the needs of a specific use case, providing helpful insights to aid the design and analysis in response to the data. In this generic example, I attempt to provide as many insights as possible while maintaining simplicity in the graph. However, the highlights can be more specific when adapting use cases with more specific needs, such as anomaly detection or key value alerts. Therefore, instead of directly visualizing a specific analysis, I add visual elements to support the extraction of multiple analyses.

I utilize absolute/global extrema to extract simple analytical highlights that are relevant to multiple domains: ranges, trends, correlations, outliers, and key moments (time steps when external events may have influenced multiple time-series) to demonstrate common analytical highlights in the summarized line graph. I plot the absolute maximum as downward pointing triangles in blue and the absolute minimum as upward triangles in red for each of the time-series. I label the triangles and align them to the mean curve using vertical dotted lines for better time point measurement and comparison. Note that the extrema themselves are not the analytical highlights, but a tool that supports easier extraction of the analytical highlights. With the absolute extrema triangles, viewers can extract the global and individual ranges using the y-values of the extrema and the approximate global and individual trends by comparing the time stamps and orders of the extrema. These characteristics provide a sufficient overview for each of the time-series data with little visual clutter (two data points each); aligning and comparing these overviews alongside the representative data allows users to identify possible correlations, outliers, and key moments and compare different subsets of the data (**R5**). Additionally, by comparing the ratio of growing trends to decaying trends, users can perform similar analyses to the market indicator of advance-decline issues. If the design instead highlights local extrema, users can also examine the local extrema to perform analyses of new highs-new lows. Finally, plotting the extrema provides measurable values of the ranges and the key moments (**R4**).

By analyzing the values and time points of the extrema, this design reduces the challenges of analyzing values of overlapping lines and lines that suffer from adjusting to the overall scale of the dataset. By extracting these characteristics from the highlighted extrema, the summarized line graph also allows users to understand how the analysis results are supported by the dataset (**R3**). This design assumes the fluctuations of the lines have a smaller vertical impact than the actual trend over time. From my observations, this is the case for most real-life data.

For example, in Fig. 4.1 where the Nasdaq stock prices of the transportation industry in the air freight/delivery service subsector during the year of 2016 are displayed, we can see that the stock price of FedEx(FDX) ranges roughly between 125 and 195 dollars and has

an overall growing trend. The figure suggests the subsector mostly shares a positive correlation with a few commodities that are curious for having neither their maxima or minima near the maxima or minima of the remaining time-series. There is, however, no obvious single outlier in the figure. The figure also suggests that the end of June is worth further exploration as six of the airline stock prices reached their absolute minimum on the same day shown by the perfectly aligned dotted lines. While this setup of analytical highlights does not target a specific scenario, these insights can prove useful to investors looking to invest in a company, managers trying to understand the performance of the company against its competitors, or security advisers searching for attacks and insider trading.

By generating these insights through interpreting the highlighted extrema, the summarized line graph also allows the audience to understand the reasoning behind the insights (**R3**). Additionally, highlighting only the absolute extrema can guarantee the number of nodes added to the graph to always be two times the number of lines in the graph rather than depending on the behavior of the dataset, and therefore keep the visual display uncluttered. The extrema are descriptive and succinct analytical highlights that enable many tasks, and additional highlights can be added while balancing the visual clutter for many applications including the example described above.

4.1.3 Data Envelope

Since the data envelope component introduced in Chapter 3.3 summarizes the remaining aggregated data, it provides important information not presented in the representative data and analytical highlights in a simplified and contextual manner. Additionally, to minimize the potential misleading information from extreme values, the data envelope adds density bands to visualize the distribution of the original lines to the visual summary. Density bands are created by placing a transparent band between the mean curve and each of the original lines. The transparency of the band is defined in equation 4.1, where C_o is the user chosen opacity, normally between 1 and 2. The equation is designed to incorporate

the standard deviation, the total range, and the line count to provide better separation of the different densities.

$$Opacity = C_o \times \frac{log(\frac{10 \times std\ dev}{max - min})}{line\ count} \tag{4.1}$$

With multiple overlapping layers, the final opacity will inform the audiences of the distributions of the lines, allowing them to better understand the original dataset and its effect on the representative data (**R3**). Density Bands also aid the audience in connecting the maximum and minimum of a line in further exploring the original dataset (**R3**), although the difficulty of such task is proportional to the complexity of the original graph. I chose not to use a conventional confidence band to preserve more information on the individual time-series. Similar to Novotný et al.'s focus+context design [135], the semi-transparency design allows the audience to focus more on the other two components. By placing the transparent bands within the 2D plane, it provides the audience with enough measurability for the data envelope component (**R4**). Additionally, by examining the density bands and the mean in the same graph, users can examine the distribution of time-series above and below the mean, similarly to the way market indicators examine the percentage of stocks above and below key moving averages.

4.1.4 Constructing the Summarized Line Graph

To construct the summarized line graph with the three components, I first scan the data in a line graph, calculate the mean value for each of the time units while keeping track of the maximum and minimum points of each line. I plot the average over time (Fig. 4.2(b)) as a summary statistics overlay [136]. I insert the nodes for each global maximum and minimum and connect the nodes to the average curve through vertical dotted lines (Fig. 4.2(c)). I add semi-transparent bands between the lines and the average curve (Fig. 4.2(d-g)). Finally, I remove the original lines (Fig. 4.2(h)) and the remaining graph is the summarized line graph utilizing the Three-Component Visual Summary design.

4.1.5 Generalizability

To demonstrate the generalizability of the Three-Component Visual Summary design, I present an alternative design using Pearson's correlation coefficient as the analytical highlights for comparing point-wise trends of individual time-series against the trend of aggregated time-series [137] for crime analytics [138]. As shown in Figure 4.3, the graph highlights the representative crime (car prowl, the time-series with a strong positive correlation; blue, with an up arrow following the label), the crime with the most opposite trend to the overall trend (narcotic, moderate negative correlation; yellow, with a down arrow following the label), and the crime most independent to the overall trend (street robbery, very weak correlation close to zero; red, with a dash following the label) alongside the average number of crime reports per month for the city of Seattle from 2008 to 2018. The line-width reflects the strength of the correlation.

4.2 Evaluation

I created a 30-minute user evaluation session for the summarized line graph design based off of the three-component summarization method. The evaluation examined the complexity and the ability of the summarized line graph to generate insights on the overall trend, outliers, and key moments against four linear time structured visualization techniques capable of communicating time-series data through static presentation: the traditional line graph, the band graph, the stream graph, and the horizon graph. Each visualization technique has its advantages and limitations for a given data set and task. Thus, I designed the tasks to be as suitable as possible to all of the visualization techniques tested.

4.2.1 Hypotheses

The ultimate design goal is for the summarized line graph to satisfy all the design requirements listed in Chapter 3.1. I tested whether this design is capable of providing a balanced and effective analysis on *all* of the tasks reflecting the design requirements, and

compare to other visualization techniques. I hypothesize the performance of the summarized line graph compared to the other visualization techniques below:

H1 *The summarized line graph will perform better in identifying outliers (improvement in accuracy and reduction in completion time) and locating key moments (reduction in completion time) compared to the traditional line graph. The two techniques will perform similarly in identifying the overall trend (improvement in accuracy and reduction in completion time).*

H2 *The summarized line graph will perform similarly to a band graph (improvement in accuracy and reduction in completion time) in identifying the original graph and the overall trend* . Because a band graph does not support examining individual time-series, the summarized line graph will perform better in identifying outliers and key moments.

H3 *The summarized line graph will perform better in identifying the original graph and outliers and locating key moments (improvement in accuracy and reduction in completion time) compared to the stream graph. The two techniques will perform similarly in identifying the overall trend (improvement in accuracy and reduction in completion time).*

H4 *The summarized line graph will perform better in identifying the original graph (improvement in accuracy and reduction in completion time) compared to the horizon graph. The two techniques will perform similarly in identifying the overall trend and outliers and locating key moments (improvement in accuracy and reduction in completion time).* The summarized line graph also supports measuring actual values which the horizon graph does not.

4.2.2 Participants

For this evaluation, I recruited 22 university student volunteers (13 male, 9 female) ranging from 18 to 32 years of age (average age of 25) with backgrounds in Computer

Science, Electrical and Computer Engineering, Industrial Engineering, Aerospace Engineering, Medicinal Chemistry, and Linguistics through the university public email lists and campus billboards. The majority of the participants were familiar with basic Excel-level visualization techniques, including the traditional line graph. There were no color-blind participants (self-reported). The participants were compensated at a $10 hourly rate. All participants were fluent in English.

4.2.3 Apparatus

The evaluation was conducted on standard Dell desktop machines equipped with a mouse, a keyboard, and a 30" monitor set to a 2560 x 1600 resolution. The evaluation was performed on a Chrome web browser page maximized on the screen. Each image was displayed at a 960 x 500 resolution. Only the mouse was used for the tasks.

4.2.4 Tasks

During the evaluation, the participants were given four type of tasks to evaluate the complexity of the visualization techniques and how they support the design requirements R1, R2, R3, and R5. Design requirement R4 was not included in the evaluation as it is straightforward from the design of the visualization techniques. The analytical tasks are inspired by Amar et al.'s taxonomy tasks [139], which explore the characteristics of an entire dataset, are not easily achievable by the majority of visualization techniques, and are reasonable for scenarios working with time-series data. I used two years of historical Nasdaq stock market data from the airline industries and four years of historical Nasdaq stock market data from the technology industries. I altered the time range and the stocks used in each question, typically a year's worth of data for 20 to 30 stocks, to prevent participants from memorizing the answer. As a result, each question was given a "unique dataset." The correct response for each task was pre-calculated using the raw data.

All considered visualization techniques were used to complete the tasks. Figure 4.4 shows how one of the tasks appeared for the different visualization techniques, although

only one visualization was given at a time, and Figure 4.5 provides examples of the four tasks and the five figures used in the study. Each task was evaluated based on completion time and correctness. I evaluated the performance of my summarized line graph on these tasks against four representative time-series visualization techniques: the traditional line graph, the band graph, the stream graph, and the horizon graph. The traditional/simple line graph, the stream/stacked graph, and the horizon graph are representative visualization techniques for displaying multiple time-series data [69], and the band graph provides a simple yet effective overview of multiple time-series data while sacrificing the ability to explore the individual series.

The traditional line graph (Fig. 4.5, top right) shares a similar visual appearance and attributes with braided graph and scattered plot [140] and is one of the most commonly used and understood visualization techniques [56]. The band graph (Fig. 4.5, left) shares similar appearance, functionality, and limitations as the river plot [141] and the functional boxplot [142] with a more simple and direct presentation. The stream graph (Fig. 4.5, top center) is a good representation of stacked graphs that highlights the overall dynamics and the individual contributors. Finally, the horizon graph (Fig. 4.5, bottom right) is a good representative visualization technique that utilizes small multiples to save space and explore both the individual and the overall dataset. I choose the horizon graph over conventional small multiples because conventional small multiples can take up noticeably more vertical space, which may not be available during the knowledge transfer between the data analysts and the decision-maker. The normalization and the different binning in the horizon graph also allow easier trend identification and comparison. The band graph, the stream graph, the horizon graph, and my summarized line graph can all be derived from the line graph. Each visualization technique was given two questions for each task. I excluded the choice "undeterminable" from the answers, forcing the participants to make their best guess when the answer is not obvious; this option complicates the calculation of accuracy and can influence the decision time measurement since the participants may give up at different levels of frustration.

Identifying the Original Graph

For each visualization technique (excluding the traditional line graph), I presented two questions per task: one for identifying the original graph composed of 20 time-series, and a similar question with 30 time-series in the graph. For possible answers, the participants were given a choice of four line graphs (each with 20 or 30 time-series, respectively) to identify as the one from which the given visualization (i.e. summarized line graph, band graph, stream band, and horizon graph) was derived. By analyzing the time and the percent of correct identifications of the original graph, I can better understand the complexity of each visualization technique and user's ability in creating the mental image of the raw form of the data through such techniques. The result reflects the visualization techniques' ability to meet the design requirement R3 (fidelity).

Identifying the Overall Trend

The participants were asked to identify the overall trend for the dataset using the five visualization techniques. For each question, the participants were asked to identify whether a given graph had an overall increasing or decreasing trend. Each visualization technique was given two questions, one with an overall growth or fall of five percent, and the other thirty percent. The result of the task reflects the visualization techniques' ability to meet the design requirement R1 (comprehensibility).

Identifying the Outlier

The participants were asked to identify the outlier using the five visualization techniques excluding the band graph. For each question, the participants were asked to select one time-series in the given graph that deviated from the overall trend the most. This task focuses on anomalous behavior, meaning a data source's value is increasing or decreasing in the opposite direction of the rest of the group, rather than a data source having values significantly higher or lower than the rest of the group. Note that I removed the band graph

starting with this task as the individual time-series are not identifiable with this visualization technique. The result of the task reflects the visualization techniques' ability to meet the design requirement R1 (comprehensibility) and R2 (accuracy).

Locating the Key Moment

Finally, the participants were asked to identify the time step when a key moment occurred (multiple time-series reaching their extrema concurrently) using the five visualization techniques excluding the band graph. For each question, the participants were asked to identify the month when the most time-series reached either their maximum or minimum concurrently in the given graph. The result of the task reflects the visualization techniques' ability to meet design requirements R2 (accuracy) and R5 (comparison).

4.2.5 Procedure

After each participant provided informed consent, I provided a 10-minute training session describing how the summarized line graph and comparative visualization techniques were derived from the traditional line graph. I then administered three sample questions, similar to the evaluation tasks, for the participants to test their understanding of the visualization techniques and the tasks to complete.

During the evaluation, the participants answered multiple-choice questions for the tasks. The evaluation question order was randomized and updated for each participant using a Latin Square randomization order [143] to ensure an even distribution of the question types throughout the evaluation trials to minimize the learning effect in the results. After the evaluation, the participants were surveyed about their demographic, self-reported skill level, and thoughts on the tasks and the visualization techniques.

Table 4.1.
T-test on the effects of difficulties.

Task	Correctness	Completion Time
Original Graph	p-value = 0.80	p-value = 0.19
Overall Trend	p-value = 0.64	p-value = 0.45

4.2.6 Results

To analyze the collected results, I examined the 95% confidence intervals calculated utilizing the bootstrapping method [144] with 1,000 iterations. Fig. 4.6 presents the accuracy and completion time of each visualization technique under each task and Fig. 4.7 presents the overall comparison between the techniques. In this section I compare the performance of the techniques using the overlap-test [145] and the t-test [146]. Note that from Table 4.1 we can see that the difference in the difficulties of the tasks to identify the original graph and the overall trend is not significant for the correctness and the completion time. Therefore, the following analysis treats the results from the different difficulties equally.

Fig. 4.6 shows the accuracy of the summarized line graph to be consistently above 80% correct. For the task to identify the original graph, summarized line graphs (μ=91%) perform significantly stronger in correctness compared to stream graphs (μ=62%) and horizon graphs (μ=33%), and similarly to band graphs (μ=93%). For the task to identify the overall trend, summarized line graphs (μ=100%) perform significantly stronger in correctness compared to stream graphs (μ=83%) and horizon graphs (μ=81%), and similarly to band graphs (μ=98%) and line graphs (μ=98%). For the task to identify the outlier, summarized line graphs (μ = 93%) perform significantly stronger in correctness compared to line graphs (μ=67%) and stream graphs (μ=22%), and similarly to horizon graphs (μ=91%). Finally, for the task to locate the key moment, summarized line graphs (μ=93%) perform significantly stronger in correctness compared to stream graphs (μ=43%) and horizon graphs (μ=71%), and similarly to line graphs (μ=79%, p-value=0.08). Fig. 4.7 shows that over the

scope of this experiment, which was designed to reflect the visualization technique's ability to satisfy the requirements listed in the introduction, the summarized line graph (μ=94%) performs significantly stronger in correctness compared to the line graph (μ=81%), the stream graph (μ=52%) and the horizon graph (μ=69%), and similarly to the band graph (μ=95%).

Fig. 4.6 and Fig. 4.7 show that the summarized line graph has the shortest average completion time in overall comparison, trend identification and key moment locating. However, the majority of the differences in the average completion time are not statistically significant. The only exception lies in the task to identify the overall trend, where the summarized line graphs (μ=23.94s) perform significantly more efficiently compared to line graphs (μ=26.98s, p-value=0.05).

The summarized line graph received positive feedback from the participants in the post-experiment survey. The participants appreciated its cleaner aesthetic and found its resemblance to the more familiar traditional line graph helpful. The participants also found the average curve and the removal of original lines useful when examining the data. Finally, the participants agreed that the visualization technique is easy to interpret and gain the insights required for the tasks. However, a few participants also expressed a minor frustration with the additional time it took to find the labels in the summarized line graph.

Comparing between the summarized line graph and the traditional line graph, I can only conclude from the study result that the summarized line graph is more accurate in identifying the outliers and more efficient in identifying the overall trend, which partially confirms and partially exceeds H1. The summarized line graph and the band graph performed roughly on the same level regarding both accuracy and efficiency, which confirms H2. However, I do note that the simple design of a band graph allows it to be a powerful tool in communicating an overview of the data, but it is not capable of the more detailed tasks a summarized line graph can handle. The summarized line graph performs significantly stronger than the stream graph in the correctness of every task, which exceeds the accuracy side of H3, but failed the completion time side of the hypothesis. Finally, the summarized line graph outperforms the horizon graph in the accuracy to identify the orig-

inal graph, to identify the overall trend, and to locate the key moment which also exceeds H4 on the correctness aspect but not the completion time aspect.

Unfortunately, due to the large variance in the completion time between the different participants, most of the comparisons between the efficiency are inconclusive. However, while the differences are not statistically significant, it shows the summarized line graph to be at least as effective as the existing techniques I tested against. I suspect the wide range of completion time is a result of some participants giving up and moving forward with random guesses at different points of time on tasks that require more effort. This theory is supported by how multiple participants expressed their frustration with the stream graph in the post-experiment survey, stating it "forced [the participants] to do a lot of the work" and was "too difficult to figure out the heights", yet the struggle is only reflected in its low accuracy but not the completion time that is similar to those of the techniques that require less effort in finding the correct answer. I consider rewarding the participants with a bonus for results above a certain level of correctness in the next study to stress the importance of getting the correct answer to the participants.

4.3 Discussion

The summarized line graph is not an intuitive visualization design and will require some training before one can use it. Based on the user study results, however, a 10-minute training is sufficient time to become reasonable skilled at understanding the visualization. Using the global maximum and minimum may be effective for examining data across a long period of time, but the audience may be confused by fluctuations when examining data that span a shorter period of time or have a stable global trend. Also, placing the labels next to the extrema makes the design less suitable for searching for specific time-series of interest without prior knowledge of their behavior. Finally, the semi-transparent density bands can also be misleading to audiences familiar with stream graphs, as the two techniques share a similar appearance but are read differently.

There are several advantages which outweigh the aforementioned drawbacks. As a shared-space technique [69], the summarized line graph's display size is independent of the number of time-series it displays, unlike techniques that create a separate chart for each series. While shared-space techniques are traditionally more efficient with fewer lines, the summarized line graph's design should reduce the impact of overlap and clutter better than traditional shared-space time-series plots as it aggregates the original lines into polygonal visual elements. Furthermore, an important advantage of my summarized line graph is that it is simple to read direct values. In comparison, horizon graphs make reading values difficult, and reading the stream graph requires estimating the width of a band. While this can be supported by interaction, such interactions are not always available.

The use of extrema and the automatic selection were chosen as "shortcuts" of typical analysis tasks, and I demonstrate the benefit of this simple design using the Three-Component Visual Summary approach. None of the participants explicitly requested additional forms of analytical highlights, and based on their performance, appeared to perform well for the specific tasks in the evaluation. However, I leave surveying domain experts on effective indicator-task combinations to future work. More complex highlights and semi-automated selection of features can be added to address the needs of other scenarios. Similarly, the scalability of the technique depends on the data and the chosen analytical highlight. From the study, the design was able to clearly visualize at least 31 time-series. However, I leave a formal scalability study to future work.

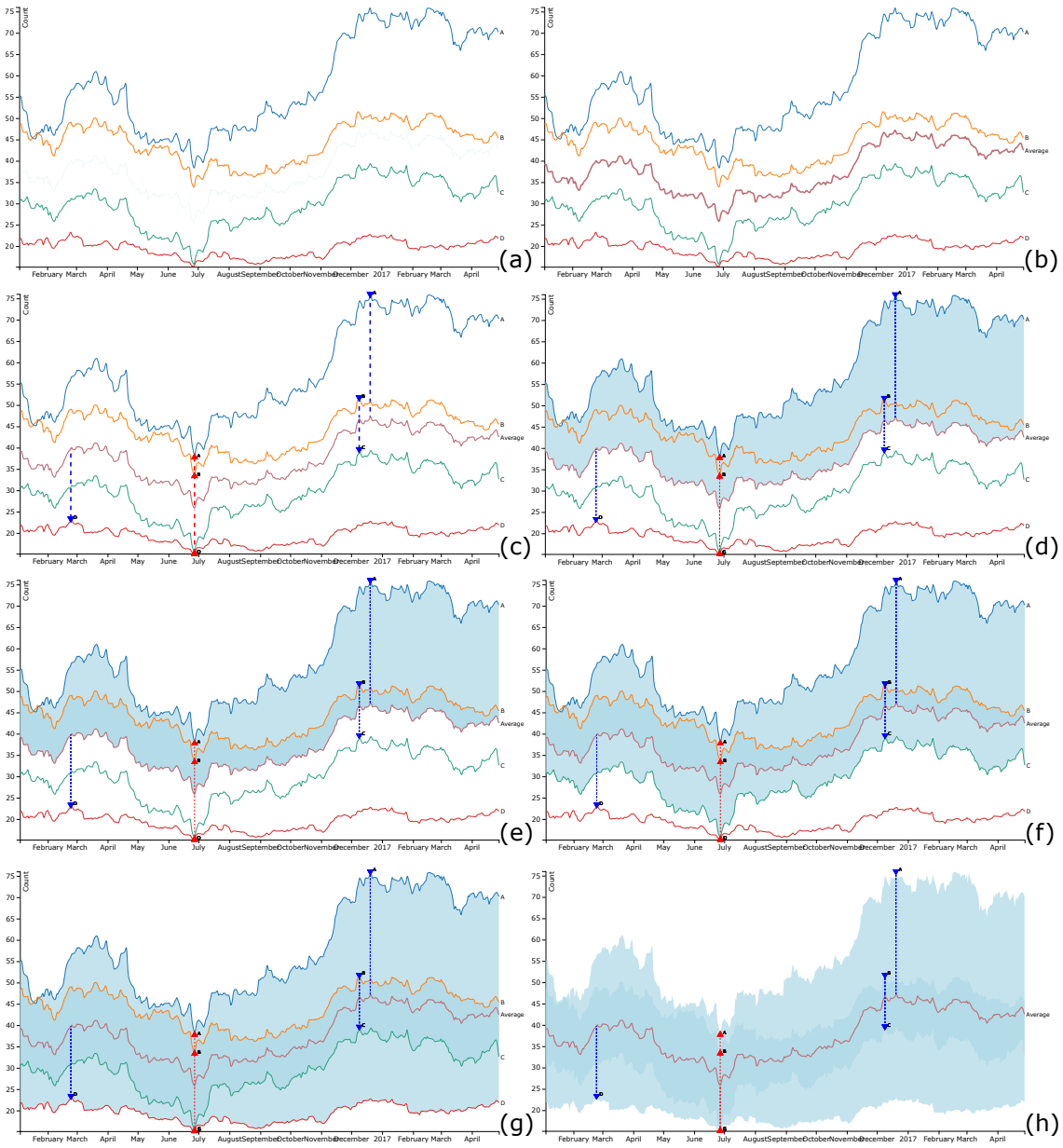


Fig. 4.2. An illustration of how a three-component summarized line graph (h) is created from a traditional line graph (a). First, the average of the original lines is plotted over time as the representative data (b). Then each absolute maximum, absolute minimum and the vertical dotted line is added to support the extraction of analytical highlights (c). Layers of transparent bands are now created between each of the lines and the average curve to form the density bands for the data envelope (d-g). Finally, the original lines are removed to reduce cluttering (h).

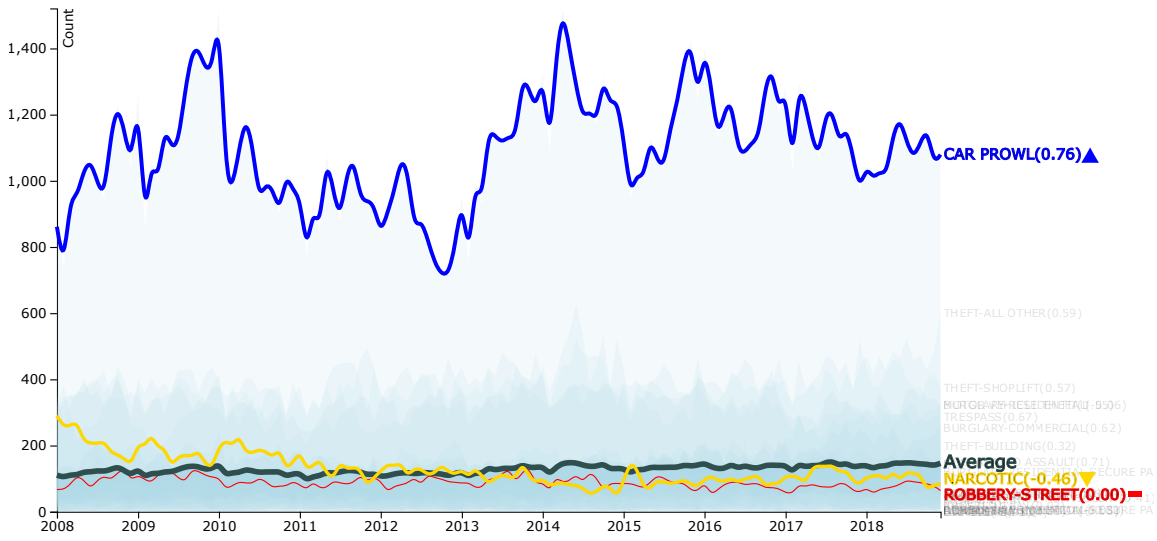


Fig. 4.3. An alternative summarized line graph design using correlation analytical highlights, showing the number of reports for 30 crime subcategories from the city of Seattle between 2008 and 2018.

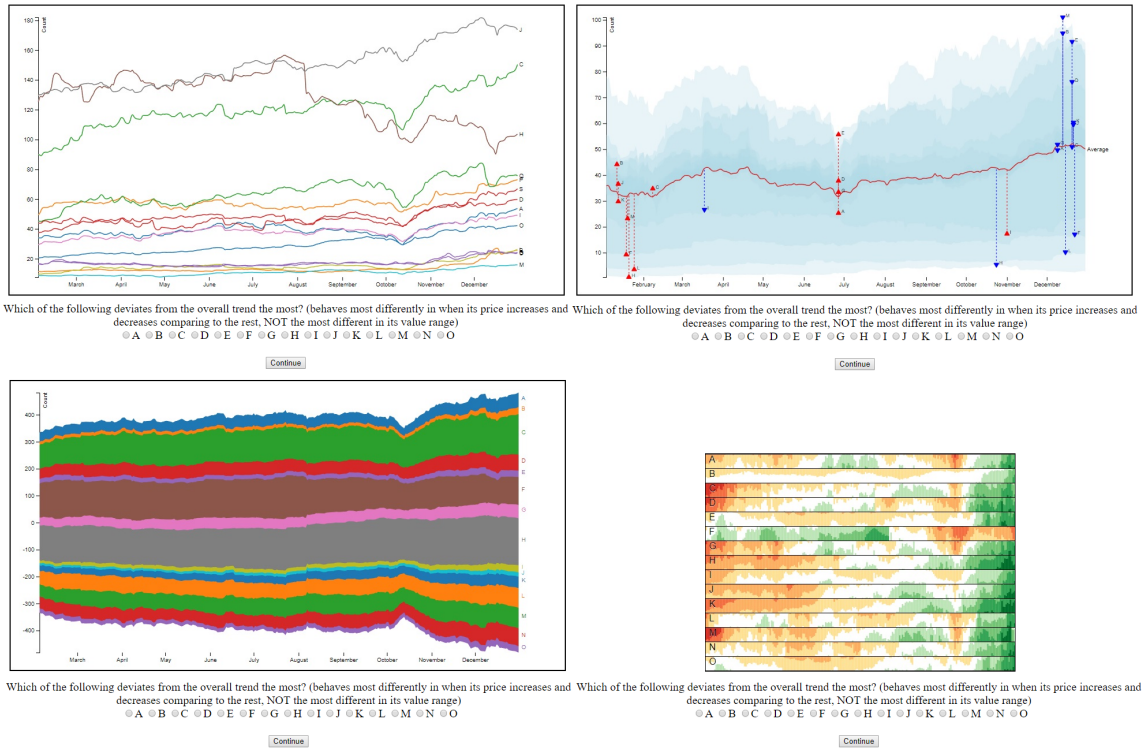


Fig. 4.4. An example of how the tasks are set up to evaluate the visualization techniques. The same task with the same difficulty (but a different dataset and answer) is given to the participants in different visualization techniques and randomized order. In this example, the participants are asked to identify the stock that deviates from the overall trend the most using the traditional line graph, the summarized line graph, the stream graph and the horizon graph. I examine the response accuracy and completion time to compare the visualization techniques.

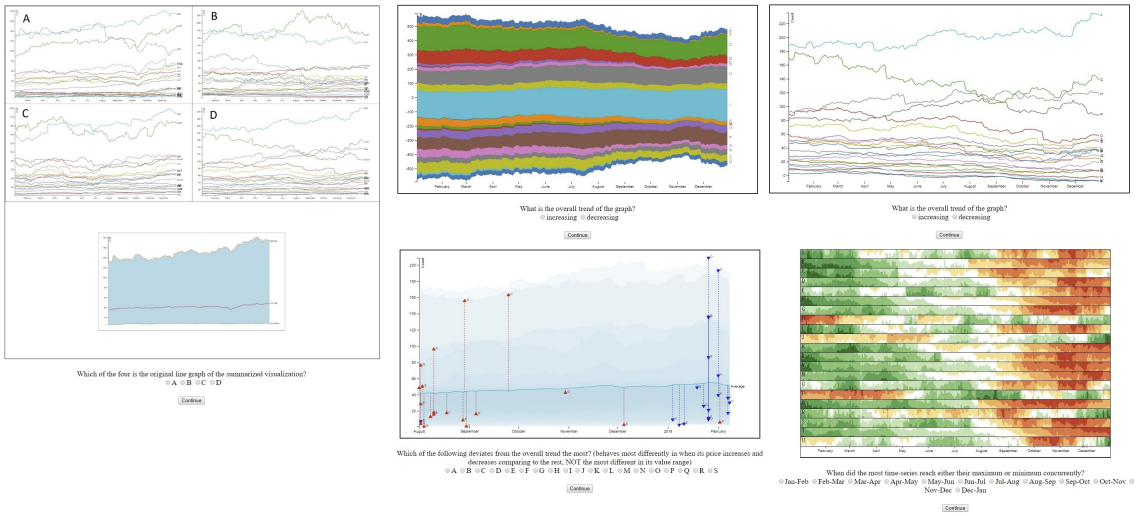


Fig. 4.5. Examples of the five visualization techniques and the four tasks used in the study: identifying the original graph using a band graph (left), identifying the overall trend using a stream graph (top center), identifying the overall trend using a traditional line graph (top right), identifying the outlier using a summarized line graph (bottom center), and locating the key moment using a horizon graph (bottom right).

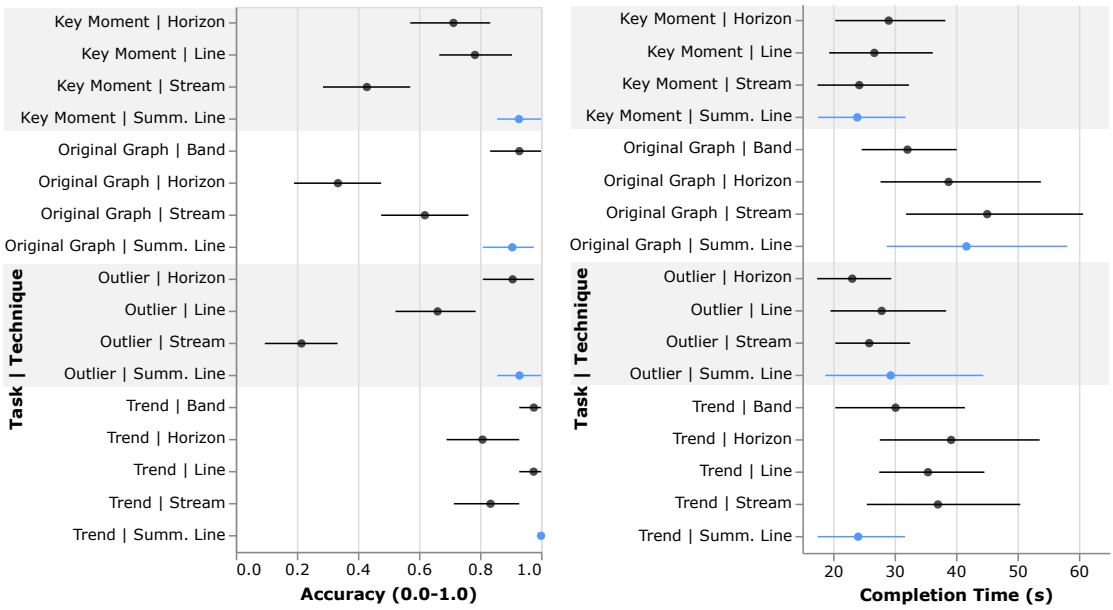


Fig. 4.6. 95% confidence interval plots of the study results in accuracy (left) and completion time (right) separated by task and technique.

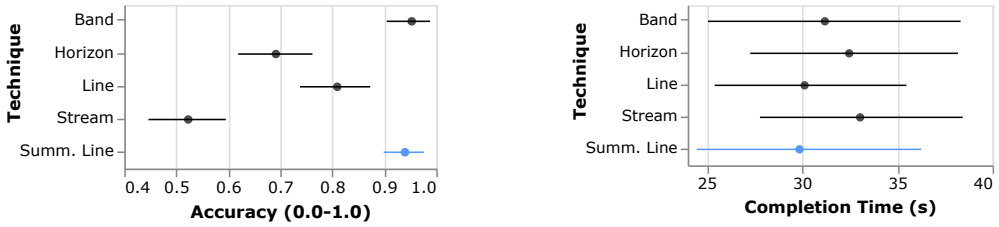


Fig. 4.7. 95% confidence interval plots of the overall comparison between the different techniques in accuracy (left) and completion time (right).

5. SUMMARIZING CONTEXTUAL DATA

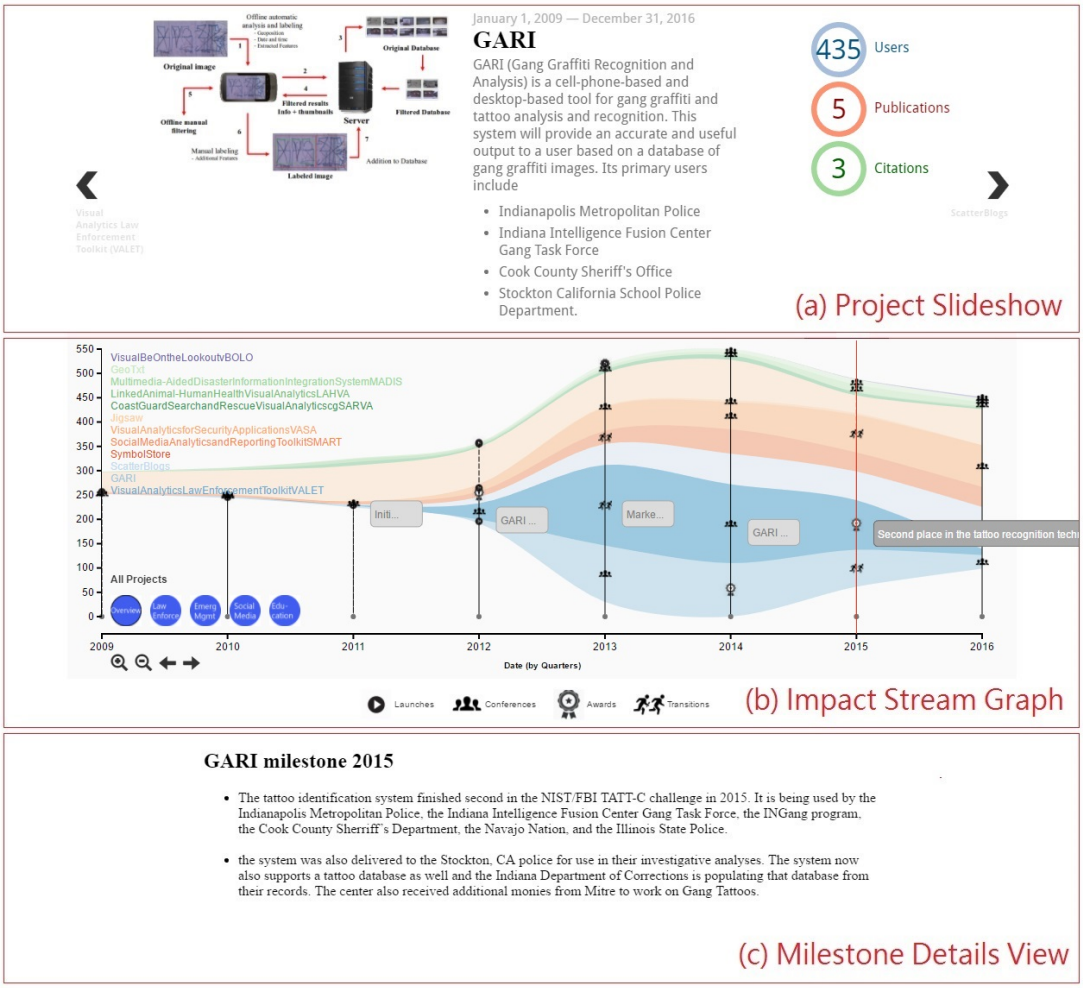


Fig. 5.1. The SuccessVis system contains three connected views. The Project Slideshow on the top (a) serves as the representative data component to provide quick summaries to the center and each project. The Impact Stream Graph in the center (b) serves as the analytical comparisons component and allows users to compare the impact of projects under the same category and the dynamics of different impact metrics within a project. The Milestone Details View on the bottom (c) serves as the data envelope component to provide context to the milestones.

A version of this chapter has been previously published in proceedings of Leipzig Symposium on Visualization in Applications 2018.
Citation: C. Yau, C. Surakitbanharn, JX. Tee, and D. S. Ebert, "SuccessVis – Visualizing Academic Impact," in *Proceedings of Leipzig Symposium on Visualization in Applications*, 2018.
URN: urn:nbn:de:bsz:15-qucosa2-328021

In this chapter, I present SuccessVis – a flexible web-based system created to present the academic impact of our research center during the eight years it was supported for its sponsors to evaluate the value of investing in academic research groups. SuccessVis was designed following the Three-Component Visual Summary design to summarize and present the knowledge embedded in a collection of text and multimedia data. Text visualization is more difficult to overlay directly on a graph compared with the numerical data design discussed in the previous chapter. Therefore, SuccessVis takes the form of a light interactive system that uses three different views as the three components.

Assessing academic impact is a field that has less support from visual analytics compared to fields such as financial management. This is likely a result of the concept of academic impact and the data required to assess impact being less defined. With our data being a collection of annual reports, visual analytics systems, presentation posters and so on, this scenario also demonstrates the challenge in visualizing the multimedia data casual experts can sometimes gather. SuccessVis addresses this challenge by grouping data into quantifiable data, contextual data, and supportive material and presenting data using components that provides insights at the different levels of exploration.

In the following sections, I will first introduce the background of this project, describe the design process and how I incorporated the Three-Component Visual Summary into the design, then discuss the resulting system.

5.1 SuccessVis

Exploring and evaluating the success of university activities and projects is challenging because of the various values in different dimensions of impact (e.g., research, education, outreach). For research activities, a common approach is to explore citation networks [72]. Researchers started visualizing the connections between citations, funding, and research topics (e.g., [74, 76, 77]) to better understand the impact of "science." As the research on academic impact evolved from the traditional "scientometrics" [147] or "informetrics" [148] and became more accessible, a new group of users, "casual experts" [77], became interested. Casual experts are often in decision-making positions and would like to support their decisions with data and facts.

During the 7^{th} year of our research center, our sponsor became interested in learning about the achievements of the center and its changes over time to help better evaluate future investments with universities.

However, most scientometric works focus on a single temporal snapshot and rarely provide insights to dynamic temporal changes (e.g., [74–77]). Moreover, the impact of academic programs can be difficult to measure compared to industry, since most outcomes come in the form of discoveries of new knowledge or developments of new techniques, rather than easily comparable and quantifiable financial profits. After discussion with the staff of the research center, I decided to develop a visual analytics system to explore and evaluate the magnitude and the temporal evolution of our success, instead of compiling a traditional text-based report.

Visualizing academic impact has its challenges. Most of the data produced by research labs are abstract and non-numerical data such as publications, presentation slides, reports, logs, etc. from different projects. Such outputs are difficult to port into traditional visualization systems [149] that are simpler and easier for casual experts to understand. To address this problem, I quantified research impact, preserved important contextual data, and then combined simple but effective visualization techniques to create an interactive exploration and analysis system of the products and projects, their impact, and their evolvement over

time. The resulting system is SuccessVis, a flexible and easily adaptable web-based visual analytics system for examining the impacts of different academic research work and how their different areas of impact change over time. SuccessVis allows users to examine and compare how impactful different projects and impact dimensions (e.g., research vs. real-world application) are and connect them to contextual data to provide a comprehensive story. The system also provides additional supportive materials on each project and milestone for users who are interested in further exploring the successes of the center.

5.2 Design Process

5.2.1 Desired Output

The first step was to decide on the desired output that would best present academic impact. An interactive visual analytics system can be a powerful tool for exploring large and complex datasets, allowing its users to obtain insights that are relevant to their decisions [150]. Being able to interactively filter and examine the center's impact in different areas visually can be more effective than examining numbers and text over spreadsheets or paragraphs when exploring the outcome of investments. However, this visual analytics system must take the skill-level of the prospective audiences into consideration. The goal, through providing a visual analytics system that utilizes constrained interactions, is to effectively communicate the scale, the temporal changes, and the context of the impact from our research center in different areas and provide an intuitive visual analytics experience for casual experts to gain insights and support decision makings.

5.2.2 Data Compilation

Determining what data was available and categorizing it was a key step in the design process. I was able to collect the majority of the data from our center universities and research came from our annual reports to the sponsor. Through the reports, I obtained information such as the publications produced, important presentations, patents received

for each project and lists of funded students from each partnering university. There were also monthly newsletters and lists of seminars and presentations from the center, as well as different presentation slides, flyers, video demonstrations, software systems and logs from the different projects. These were considered non-numerical contextual data.

I needed to separate the data into quantifiable data, contextual data, and supportive materials in order to visualize the available contextual data using more commonly understandable techniques that typically require numerical data. Quantifiable data are variables whose significance and magnitude are both positively correlated and can be represented in or transformed into a countable manner. For example, the amount of funding a center obtains may be proportional to the outside interest in its research work. Contextual data includes variables whose significance are not directly represented by their count. A prestigious award might be a more significant outcome than multiple less prestigious awards, for example. Finally, supportive materials, such as a recorded demonstration of a working system or the paper from a research discovery, provide additional details into the story for the impact variables.

I acquired the number of publications per year for each project, and then searched for the number of times these papers were cited throughout the years of interest from the list of publications. I then estimated the number of important presentations from our annual reports and quarterly newsletters. I either obtained the exact number of users based on the accounts created, or estimated the number of users based on the unique download IP addresses from the system logs. These classes of data became numerical to support more traditional visualization techniques. The data was manually converted into counts during this process, as the source materials were stored in different mediums and were accessed differently. I leave the automated conversion to future work. I also collected contextual data such as project milestones, start dates and end dates, significant project transitions, workshops hosted, competitions participated, and awards received from annual reports and newsletters. The contextual information would be presented as textual data and connected to the visual components by date and project. Finally, there were additional supportive

materials such as posters, presentation slides, and videos that could be attached to relevant textual data for further investigations.

Since I aimed to present multiple projects from the center, I needed to find the data fields commonly shared, and therefore comparable, that could represent the scale of impact. After examining the amount of data from each data field for our main projects, I finalized four fields or areas of impact: publication counts, citation counts, presentation counts, and user counts. All four fields, besides user counts on rare occasions, are fairly common across the different types of academic projects and cover a variety of impact types. The number of publications reflects the novelty and depth of the research work, the number of citations reflects the relevance to other research works, the number of presentations reflects the interest from the outside world, and the number of users reflects the practicality of the research output.

5.2.3 Visualization – the Three-Component Visual Summary

To effectively present the center impact to the stakeholders, I constructed a Three-Component Visual Summary design with the data collected. Unlike the summarized line graph that combines multiple numerical data visual representations, it is difficult to overlay visual representations of the contextual data and supportive materials. This is because there is no commonly shared basis to connect the different components visually onto one canvas. In this use case, I display the three components in three interactive and interconnected views with the view that presents the representative data on the top and the view that presents the data envelope on the bottom. The information to present in the three different views was identified based on conversations with the stakeholders. The subsections below describe the component designs for the representative data, the analytical comparisons, and the data envelope.

Representative Data

The representative data component provides users with an overview of the dataset. This component must cover both the quantified portion and the contextual portion of the dataset to guarantee the delivery of a comprehensive summary. This component also needs to provide a summary to both the center and the primary sources of the impact – the major projects from the center. Therefore, a Project Slideshow View was created as the representative data component. The project slideshow presents a high-level description and the total count of each impact metric for the center and the various important projects to inform users of the scope and direction of the impact generated by the center and by each project. The project slideshow also updates automatically based on the data being examined in the analytical comparisons component and highlights the corresponding story to help users maintain perspective and ensure users to connect the three components in the three different views.

Analytical Comparisons

The analytical comparisons component allows decision-makers to perform comparative analyses and quickly generate insights useful to the decisions. This component presents the extracted and quantified high-level attributes shared between data subsets to allow accurate and effective comparisons. The Impact Stream Graph was implemented as the analytical comparisons component to provide audiences the opportunity for comparative analysis on the magnitude and temporal evolution of the different impact metrics of a project or the aggregated impact of different projects. By plotting the quantified magnitude of impact over time, this visualization allows comparisons that are more intuitive than text-based reports.

The Impact Stream Graph was implemented with a stacked graph using the ThemeRiver layout [37] to visualize the scale of impact and its changes over time. Alternative visualization techniques that can encode magnitude over time for multiple distinguishable time-series include the traditional line graph [20] and the horizon graph [67]. However,

the layout of the traditional line graph does not reflect the combined magnitude and the proportion of the individual time-series, and the normalized horizon graph makes comparing the magnitude of each time-series impossible. The strength of the ThemeRiver layout lies in visualizing thematic variations in both individual topics and groups of topics over time [47], which fits the need to show the impact of the center as a whole, the impact of individual projects, and the different types of impact from each project in a measurable and comparable manner. Being able to see how the different impact metrics of the different project categories evolve over time allows investors to evaluate the direction and timespan of their next investment. ThemeRiver is also constructed on the same abstract 2D plan as the traditional line graph and therefore reduces the learning curve for casual experts to use effectively compared to designs like the horizon graph.

Icons were added to the ThemeRiver to represent milestones whose significance cannot be represented fairly by its count. Different glyphs and simple annotations provide users with a quick look into the significance of the milestones and how they might be pivotal moments to the impact generated. A click event on the icons will update the data envelope component to provide more details on demand.

Data Envelope

The data envelope component provides context embedded in the raw data to the first two components. With this dataset, the component needs to be able to incorporate the different data formats of the raw data. The Milestone Details View serves as the data envelope component and provides the audience with a chance to look into the details of the different milestones. This view provides more detailed descriptions of the milestone stories highlighted in the Impact Stream Graph as more digestible aggregated summaries, and provides links to the relevant original materials to incorporate the diverse formats of the raw data.

5.2.4 User-Centered Design

I designed the system with an iterative and incremental development approach [151]. I first presented a mock-up to our sponsor, a decision-maker in the Department of Homeland Security. I created a prototype system using their feedback. After the second development iteration and collection of feedback using the prototype, I refined the design and populated it with three major projects from our research center. With the affirmation on the prototype, I populated the full data and continued demonstrating the refined system to our sponsor and then incorporating their feedback back into the system. To ensure the design appeals to casual experts, I also reached out to an Interaction Design & Industrial Design team at the university to improve the color choices and glyph designs for the milestone icons used in the system.

I distributed the work of collecting and filtering the milestone contextual data to administrative staff, instead of the researchers, to minimize the possibility of bias regarding what would be considered as important stories to tell. For example, researchers see the discovery and overcoming of roadblocks during a project as important achievements, while sponsors pay more attention to project launches, transitions, and awards. It is important that the information the system provides is relevant to the audience rather than the presenter.

Finally, to make the system usable over time as center activities continue, and to make it beneficial to other centers without redevelopment, I decided to create a system that is simple to update and flexible to adapt to different fields and measures. I also decided to make the system web-based to avoid installation issues and enable easy access from anywhere for use and presentation. The system stores data and populates the user interface through spreadsheets for the ease in updating and exporting from databases, etc.

5.3 System

Figure 5.1 displays the system consisting of three parts: (1) a Project Slideshow (Figure 5.1a) that gives summary highlights for each project, (2) an Impact Stream Graph (Figure 5.1b) that displays the magnitude of the impact and its changes over time for the different

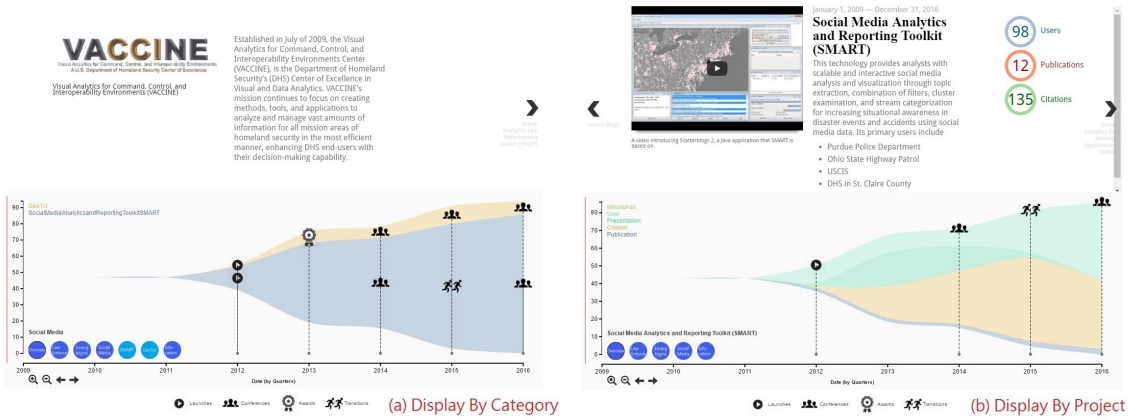


Fig. 5.2. SuccessVis displaying the impact breakdown of (a) the Social Media Category and (b) the project SMART.

projects or types of impact, and (3) a Milestone Details View (Figure 5.1c) that allows users to read the milestone story in more detail and access the attached supportive materials.

5.3.1 Data Spreadsheet

To populate the system, a few specific spreadsheets have to be generated:

- "slideshow.csv" stores the Project Slideshow information. It includes, for each project, its project name, start date, end date, project description, and an additional media link.
- "gauges.csv" stores the subjects and the values of the three takeaway values on the right side of the Project Slideshow.
- "metrics.csv" stores an index number and a corresponding name for each tab the Impact Stream Graph displays.
- "visualization_data_x.csv", where 'x' is a positive integer, stores the actual data of a tab in the Impact Stream Graph. Its fields include a topic key, a topic value, a date, a milestone summary, a milestone glyph, and a milestone link.

- "milestone_legend.csv" stores the paths to the milestone glyph image files and their corresponding milestone categories, then uses the data to generate the legend at the bottom of the Impact Stream Graph.

5.3.2 Visual Analytics

The three views of SuccessVis (Project Slideshow View, Impact Stream Graph, and Milestone Details View) combine and link the numerical data, the contextual data, and the supportive materials to provide a full picture of the center's impact.

Project Slideshow View

Project Slideshow View includes three subcomponents: project description, external media, and takeaway values. The project description, displayed in the middle of the Project Slideshow, provides a summary of the purpose, the output, and the partners of the project. The external media window, displayed on the left side of the Project Slideshow, is capable of displaying an image, a video, or a PDF file as long as a valid link address is stored in the spreadsheet. The takeaway values are located on the right side of the Project Slideshow. It provides users with values on the project as a whole. The slideshow starts off with an initial slide on the center itself, then lists the projects ordered from left to right in chronological order. Users can click on the left or the right arrows to navigate through the projects. A semi-transparent summary that includes the project name and timespan of the next/previous project, located under the arrows, will be highlighted when users hover the mouse over the arrows. The project the slideshow is displaying will also be highlighted on the Impact Stream Graph if it is included in the selected tab.

5.3.3 Impact Stream Graph

Impact Stream Graph displays the impact of the center's work. It is capable of displaying multiple tabs of impact, which are selected through the blue circle icons in the lower

left corner, allowing users to display and examine impacts from different projects and categories. Tab one displays all the projects' collective impact, as seen in Figure 5.1b, showing how the impact from the center as a whole has evolved. The substreams represent the total impact of each project showing the proportion of their contributions to the overall impact of the center. While examining our projects, I grouped the projects into different categories such as law enforcement, resource allocation, and social media. In Figure 5.2a a user has selected the Social Media tab, leading the Impact Stream Graph to display the collective impact of GeoTxt and SMART from the Social Media category and the proportion they each contributed to the combined magnitude. By double clicking on the larger substream that represents SMART or by clicking the SMART tab icon, the graph will display a project impact breakdown with the substreams being the different impact fields, as shown in Figure 5.2b, allowing the user to better understand what areas the projects have a stronger impact on. Different icons are placed on top of the stream graph at the corresponding time to show important milestones each project achieved. The different glyphs represent different categories of milestones such as project launching, transitions, and awards. By hovering over a milestone icon, a summary of the milestone story will be displayed. By clicking on the milestone icon, a detailed description of the milestone will be displayed in the Milestone Details View, as shown in Figure 5.1c. When a project is highlighted, either through the slideshow or through hovering the mouse over the substream, the milestone icons belonging to the project will be highlighted to help users better connect the projects and the milestones.

5.3.4 Milestone Details

Milestone Details View provides a more in-depth description of milestones. It also displays other important events that happened during the same timeframe that are unable to be displayed on the milestone summary in the Impact Stream Graph. Milestone Details View takes a web URL link and displays it in an iframe when its corresponding milestone icon is clicked. By displaying a webpage, it is capable of not only explaining the milestone stories

in more detail, but also embedding images, videos, external links, or other attachment files for users to examine.

5.4 Discussion

5.4.1 Use Case

I used SuccessVis to visualize the impact of our research center from the past eight years. I selected 11 major projects to display in the system to learn about the strengths and weaknesses of the different projects and project categories regarding impact. For example, by examining the law enforcement category, we can see three of the four projects have more users compared to most of the projects, indicating that the law enforcement projects have a strong impact in the work field. I can also examine how the proportions of different impact fields from a project changed as time progress. For example, in Figure 5.2(b), we can see how around the year 2015, SMART's users started increasing more dramatically while the number of presentations and citations started decreasing, indicating the overall impact transitioned from a more theoretical interest into more practical uses after about three years of development.

5.4.2 Generalizability

The system has been designed to be adaptable to other academic centers. Three out of the four impact types are collected from common products of most academic research centers. The user count and other possible impact factors not included can be easily added or removed using the populate-through-spreadsheet method. While the system does not currently support automated source material conversion directly, the process can be replicated manually and is easily semi-automated through periodic data collection, rather than gathered at the end of the period of interest.

5.4.3 Challenges

In spite of the initial system, there remain challenges to be solved, such as:

- Data frequency: Paper publications and citations often happen at the rate of one every few months, while new users and presentations vary between the different projects. If I input the stream graph data by month, I could end up with multiple zeros, which creates unattractive visuals and may not present the impact of the publications and citations fairly. However, when I input the data by year, I can only display one milestone icon per project per year due to the limitation of the spreadsheet-filled system. Multiple milestones of different kinds could occur within the same year, and it is difficult to represent the different events with just one icon.

- Limited usable data fields: To ensure reasonable comparison between the impact magnitudes of different projects, the project impacts have to share the same quantifiable data fields. I am thus limited to only the data that exists for all the projects I plan to display.

- Combining impacts: To compare the impacts from different projects, I need a representative overall impact value for each project. To not be biased toward a specific field, I add up the four different fields of impact to get an estimated project impact. However, the four fields do not share the same units or weights, and adding up the fields makes the actual measurement confusing. If I instead normalize the different fields before adding them up, the comparison between the project magnitudes will not be accurate.

- Project duration: Not all the projects share a similar amount of data. A simple spreadsheet may be more effective for short projects than the system. But selecting a threshold at which the system will be most effective is difficult.

5.4.4 Initial Feedback

After presentations of the system using eight years' worth of data, our sponsor indicated an appreciation for the storytelling aspects of the system and the ease of understanding and comparing the overall picture of the center and its projects. Our sponsor was very interested in sharing the system and in expanding the use of the system to university research centers to show their impact as well. Our sponsor also expressed interest in learning more about the educational impacts. To respond to this request, I collected data regarding courses and students that benefited from the funding and inserted it into the system as an additional and standalone project.

I had the chance to present this work to faculty from other universities. Many of them expressed interest in having access to the system and the opportunity to populate their research impact. I hope to collaborate with them in the future to further evaluate and improve the system.

One piece of feedback expressed concern about a potential color matching issue between the takeaway values in the Project Slideshow and the impact type in the Impact Stream Graph. This is, unfortunately, a result of the approach to populating the system through dynamic spreadsheets. Linking the repeated fields between the takeaway values in the Project Sideshow and the fields in the Impact Stream Graph is currently not supported. I aim to resolve this issue in future work.

These pieces of feedback were collected over multiple demonstrations and informal interviews with the intended user of the system and university faculties that can benefit from the system. No formal reports were generated.

5.4.5 The Success of SuccessVis

Unlike the summarized line graph, it is more challenging to perform a quantitative study on the effectiveness of SuccessVis due to the more interactive nature of the design and the lack of well-developed tools focusing on the exploration of academic impact over time. I also do not have access to information on how future decisions in academic-based

investments have benefited from SuccessVis. However, the value of this work is reflected in the following: First, I have successfully created and populated a visual analytics tool based on the Three-Component Visual Summary design using the type of unorganized data collection many academic research groups have, which, in the end, is preferred over the traditional reports. Second, since our sponsors wish to provide the system to other academic research centers and other university faculty members wish to adapt the system to showcase their work, it is clear that SuccessVis is a desirable storytelling tool for academic impact. Therefore, even though SuccessVis, as a Three-Component Visual Summary design, is not evaluated over a quantitative study, it is evident that it outputs an effective summary presentation which allows the casual experts to evaluate the impact of the VACCINE Center.

6. SUMMARIZING GEOSPATIAL DATA

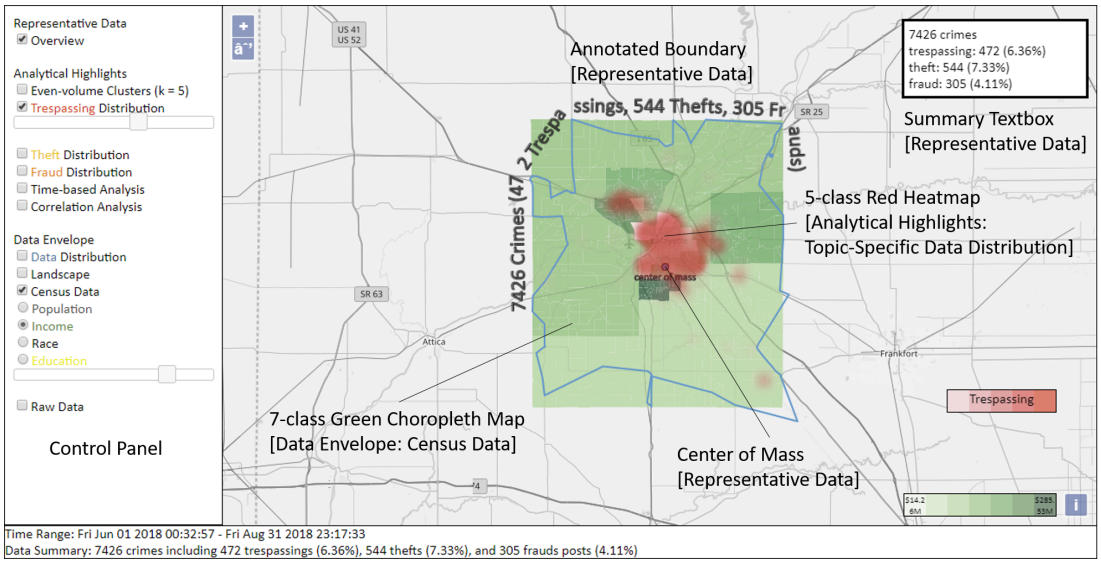


Fig. 6.1. A web-based Three-Component Visual Summary design for geospatial crime report data. Users can select and adjust the appropriate visual elements for the three components through the control panel on screen left. This design utilizes an annotated boundary and a summary textbox as the representative data to provide a quick overview of the dataset. Users can select a combination of even-volume clusters, topic-specific data distribution(s), time-based analysis, and correlation analysis as the analytical highlight for decision-relevant insights. Users can also select a combination of the raw data distribution, landscape information, and census data as the data envelope for context.

Large amounts of multimedia and multivariate data are now geo-tagged by the GPS trackers embedded in smartphones and other devices [152]. This presents new opportunities for data analysts and stakeholders to explore and identify patterns and relationships between the spatial distributions and various other spatial characteristics over different subjects of interest. The connection to geographic space also provides a unique opportunity to incorporate knowledge of the geographic region into the decision-making process. Previous

applications included in this dissertation render data in an abstract 2D space disconnected from the physical reality. Incorporating knowledge of different geospatial regions implies that the synergy of human and computer is specifically beneficial in this use case. However, the current tools and methods for analyzing geospatial data prove challenging for stakeholders and decision-makers from a more diverse background who are not trained in data analysis to use effectively [88]. This results in a more challenging collaboration between different actors, such as data analysts, decision-makers, and stakeholders. Often, presentations with filtered options from data analysts are not enough for decision-makers to make data-driven decisions confidently [153]. In the case of geo-tagged historical reports, visual analytics tools were developed to support decision-makers in the first responder fields to identify potential risks following the different decisions [44, 45]. With additional knowledge of the geographic locations, decision-makers are also able to consider factors that are not encoded into the data, including the atmosphere of the various neighborhoods and how communities may respond to the decisions, etc. However, it is observed and confirmed by our domain experts that and many of the decisions are still made primarily based on domain knowledge and traditional methods. Further, first responders primarily operate on a case by case scenario and often miss opportunities provided by the collected data.

In this chapter, I propose a web-based visual analytics system to support decision-makers in first responder fields in utilizing multivariate geospatial data more effectively for situation assessment and resource allocation. This design follows the Three-Component Visual Summary design to address the main issues in Geovisual Analytics for Spatial Decision Support: collaboration, communication, and flexibility [153]. The Three-Component Visual Summary design provides the solution by being: (a) accessible to casual experts, (b) capable of generating comprehensive visual summaries, and (c) customizable for specific decisions. It presents a dataset using three simultaneously displayed visual components (*representative data*, *analytical highlights*, and *data envelope*) encoding the knowledge generally retrieved at the high-level overview, the comparative analysis, and the low-level detail stage of exploration into one display. The three-component approach communicates a comprehensive data story more efficiently without requiring the technical expertise

to interactively explore a dataset using a visual analytics system, and the displayed visual components can be customized to focus on analysis more relevant to its audiences. The proposed design, as shown in Fig. 6.1, utilizes an annotated contour and a summary textbox as the representative data, the different possible combinations of data clusters, topic-specific data distribution, data movement, topic-specific correlation as the analytical highlights, and a density function visualization of the entire dataset, landscape, and census data as the data envelope. While the data envelope component is traditionally limited to summarizing the aggregated data, this design imports external contexts that are not encoded in the data entries but have the potential to explain the patterns observed, similarly to the domain knowledge of a user. Simple human-interface interactions are incorporated into the system to provide the audiences with the ability to explore the dataset and the analyses further and adjust the transparency of each visual element to generate a balanced visual summary.

The following sections explain the design of the system in more detail, present interesting findings through a case study and the feedback from a prospective user, and discuss the outcome of the design.

6.1 Design

In this section, I first explain the design choice behind each of the three components. I then describe the system setup. The three components were built over a map visualization for two reasons. First, most spatial analysis can be visualized over geographical coordinate systems, meaning the three components can be overlaid directly to allow are more straightforward mental connection. Second, with most geospatial data visualized using maps [154], this setup will be more familiar to casual experts. Both the direct connection between the components and the familiar setup aim to create a more accessible tool for different actors. This design considers data-driven, political (man-made), and geographic (natural) boundaries to present a more coherent picture of the region of interest.

6.1.1 Representative Data

The representative data component provides a quick summary to help users understand the scope of the dataset. The visual design has to be simple, effective, and prominent to attract the user's attention upon first glance. Therefore, a simple contour is used to highlight the boundary of the dataset. The contour boundary is quick to extract, requires no interpretation, and informs users the geographical range of the data points in the dataset immediately. It encloses the remaining visual components, establishing its role as the first element to examine. Its hollow nature also minimizes possible collisions with other visual components. However, the boundary itself provides no information as to the size and distribution of the data, which can be important in order for users to understand the scope of the dataset. Therefore, the dataset's center of mass is visualized and labeled on the map to hint at the data distribution and provide a representative point for the dataset I annotate the boundary by adding a short text description on the total count of the data and the number of data points that fall under specific categories relevant to the decision-makers to the outer edge of the contour. Another challenge with the boundary overview is that it may not be visible in the display when a user zooms in on a region within. While the remaining visual components can inform the user whether the displayed region contains data entries, the data summary contained in the annotated text will be lost. An easily accessible fixed-position textbox is added to the corner of the map to provide the summary in more detail.

A few alternative options to address the challenge in examining the overview while zoomed in include using a fisheye [155] or a space folding [29] distortion and an additional zoomed out map. However, maintaining the true scale may be important to some users, especially in the static visual output for communication. With visual components for analytical highlights and the data envelope also present, the distortion could bring more harm than benefit. The additional map view also moves the overview components out of the same coordinate scale, making the connection between different components more difficult. The additional textbox ensures a smoother exploration of the three components.

6.1.2 Analytical Highlights

The analytical highlights component presents interesting characteristics of geospatial data that can be combined in the visual display for further exploration. A list of geospatial analyses that are beneficial for data-driven situation assessment and resource allocation in first responder fields is added to the design as the analytical highlights component. The analysis results are encoded with visual components that can be understood by most audiences to allow more fluent communication and collaboration.

First, the design provides even-volume clusters. Traditionally, first responder resources are allocated to cover evenly-sized regions. However, different regions may have varied amounts of incidents happening. By allocating the resources based on even-volume regions, first responders can guarantee having enough resources for each region. The even-volume clusters aggregate data points into multiple clusters, each with the same number of data entries, using a Same-size k-Means Variation algorithm [1]. Each cluster is visualized using a contour visualization. The contour visualization is selected to allow further examination of the regions within. Users can layer the contour over data distributions to understand the hotspots within the region. Users are also given the ability to adjust the k-value through the corresponding slider bar in the control panel. This option allows decision-makers to adjust the number of clusters based on the resources available. A logical next step would be to automatically adjust the contour to conform to the shape of the streets or the neighborhoods if available. I leave that to future work.

The topic-specific data distributions visualize through heatmaps the distributions of data points that fall under a specific category or include specific content. Decision-makers in law enforcement have started considering assigning officers based on their expertise for more effective work [156]. Understanding the distributions of different incidents allows decision-makers to assign the appropriate resources to each region. A heatmap visualization is chosen, as the density may be more crucial than the range when the resources are ranked by relevance. The topic-specific distribution option also allows users to examine how this

[1] https://elki-project.github.io/tutorial/same-size_k_means

subgroup of data contributes to the overall dataset and how it relates to the political and geographical boundaries when layered over the corresponding data envelope components. Users can select the topic of interest using the corresponding radio buttons in the control panel. To ensure the system complies with the recommended constrained interaction, the system does not allow users to filter the topic through queries but has users select from a pre-determined list of topics. If the relevant topic or category is not determined, a dropdown menu may suffice. In this example, we limit the number of topics to reduce the complexity of the system.

Time-based analysis encodes the directions and the incident counts of a selected group of data over multiple time bins during the selected time frame. Understanding potential seasonal or hourly patterns of incidents over space and time can support first responders with situation assessment and resource allocation [44]. Users can filter the data displayed by topic, select the number of bins to separate the data into, and alter the start and end time of the data displayed using the radio button, the number-only input field, and the double slider bar in the control panel. This allows users to potentially display the data movement by the hour, day, week, month, year, etc. and identify possible temporal patterns in the changes of data size and weighted center over time. The data movement is visualized using circles and arrows. The position of the circles represents the data collection's centers of mass for each time bin. The radius of the circles represents the size of the data collection for each time bin. Finally, the arrows connect the circles pointing from the circle of the earliest time bin to the circle of the latest time bin. Additional contour boundaries are added to present the geographical range of the data collections' distributions to address the potential confusing from the radius of the circles. The contours can be distinguished by the opacity where the lighter contours belong to earlier time bins, and the darker contours the later time bins. This visual encoding is chosen to focus on the high-level shifting (or overlapping) of the aggregated data, which, unlike trajectory-based movement visualization, is not bounded by street paths. An alternative visual encoding uses multiple heatmaps for the different time bins and indicates the order by a sequential color scheme. However, while the heatmap encoding is more effective in showing the range of the data collections,

the movement can be more challenging to track, the normalized visualization is less effective for comparing data sizes (which should not be confused with the geographical range of the data distribution), and the area-covering visualization style can create more visual clutter when layering multiple visual components. The currently-visualized time frame is displayed in the summary textbox to reduce the visual clutter in the map view.

Correlation analysis encodes the correlation between the data subsets of two categories within the different political boundaries. This allows first responders to understand the relationship between the two categories (e.g., criminal charges) in different geographical regions better [52] and can support them in situation assessment and making potential predictions based on the data collected on one of the categories. The correlation is visualized using a choropleth map that utilizes census tract boundaries with a diverging color scheme. An alternative design could use a grid-based boundary and encode the statistical significance of the correlation on the side [52]. However, census data were assigned to specific census tracts, and by visualizing along the census tracts, the design retains the visualization of the base map and reduce the hindrance from examining the details in the base map. While the nonuniform boundary in this design makes incorporating indications of statistical significance difficult, the statistical significance can be estimated by layering the corresponding data distributions. Users can select from the control panel the category pair of interest. The system then calculates the correlation for each census tract using the phi coefficient [157] and colors the corresponding census tract using the computed value. The calculated phi coefficient will have a value between negative one and one, where negative one represents the two categories strongly negatively correlated, zero represents the two categories not correlated, and one represents categories strongly positively correlated. The diverging color scheme was chosen as being able to identify correlations close to zero is equally important as being able to identify coloration near one or negative one.

6.1.3 Data Envelope

The data envelope component summarizes the remaining data and provides context to the analytical highlights. Unlike the numerical and the contextual applications that draw contexts from the raw data only, geospatial data can also retrieve context from additional geographical and political information on the region the data is geotagged. As a result, the data envelope component of the geospatial application includes the aggregation of the original dataset, landscape, and census data.

First, the data envelope summarizes the dataset by visualizing the distribution of all the data points using a heatmap visualization. The density-based visualization allows a well-scalable presentation of the raw data distribution. With the proper opacity level, it can overlap with analytical highlights to inform users how the data distribution could have contributed to the analysis results retrieved or how a highlight compares to the overall dataset in volume, etc. This heatmap uses a color scheme in the opposite color family to the topic-specific data distribution heatmap visualizations to allow more distinguishable comparisons.

Another method of visualizing the raw dataset is to directly display the geo-tagged data entries as individual points on the map. However, this introduces visual clutter quickly with larger datasets. The circle visual components can easily collide with other visual components, and data entries generated from the same geographical coordinates will overlap perfectly and become unidentifiable. This visualization may be more useful when not displayed with additional visual components, but for merely examining the raw data. The system keeps this visualization as an option for flexibility as the point visualization may be more intuitive compared to heatmaps for certain casual experts. Data entries that belong to the highlighted categories are filled with the same colors used in the topic-specific visualizations.

The landscape option updates the base map to include landscape information. This additional context can be crucial as certain phenomena are caused by landscapes [94], while the landscape data is rarely encoded in the raw data. This allows users to examine

the relationship between the terrain and the data distributions or movements. The landscape information has the potential to provide reasoning to various analytical highlights.

Finally, the census data is added to the visualization using a choropleth map. Census data provides valuable characteristics of the different neighborhoods that are not likely included in the raw data. Domain experts familiar with the regions examined may be able to provide similar information that is more up to date. However, the census data can be more detailed and be encoded into the visualization for more precise comparisons. Users can choose to visualize population, income, race, or education data using the radio buttons in the control panel. Hovering over each census tract in the map will also trigger a short text summary of the actual count inside the selected census tract to appear in a summary textbox below the map view. Both population and income data color the census tracts using a scale of normalized range from the lowest to the highest value. Education data colors the census tracts using a scale from no education to Ph.D. Race data colors the census tracts by mixing white, black, yellow, and red based on the proportion of Caucasian Americans (white), African Americans (black), Asian Americans (yellow), and Native Americans (red) residing in each census tract. By overlapping the census data with visual elements from the analytical highlights, users can explore potential reasoning or correlations behind how the different characteristics change over the different regions.

6.1.4 System

The geospatial application of the Three-Component Visual Summary design takes the form of a lightweight web-based visual analytics system that outputs static visual summaries by selecting and layering the different combinations of visual elements. The system was implemented in a web-based environment to reduce the technical requirements associated with software installation and allow accessibility through more devices and platforms. Being able to display the system in a touchscreen tablet also allows more opportunities in presenting the data. This fits the overall direction of the Three-Component Visual Summary design to be more accessible to users.

The system interface utilizes a map view, a control panel, and a summary textbox. The control panel allows users to adjust the visual presentation for each of the components in the system. Each of the visual components can be displayed or hidden using a checkbox. When displayed, each option (asides from the overview and the landscape functions) is given a slider to adjust its transparency for users to find a balance between the layered visualizations displayed for the most effective storytelling presentation. A tooltip will appear to provide a short description for each option when hovering over a label or a slider. Finally, the summary textbox displays text summaries of the overall dataset and additional regional information when the appropriate visual elements are selected to be displayed.

While the visualization relies on the "three components" to tell the data story, all three components visualize the data over the same coordinate system and therefore are designed to directly overlay on top of each other to preserve space and provide stronger mental connections between the components. Each of the visualization techniques is also selected and implemented to be compatible, meaning they should be distinguishable and understandable when displayed together.

This system and approach combine the computational power of machines and the soft knowledge of the user. The machine computes and visualizes the dataset, and given the ability to adjust the visibility of each visual component, the end users can use their domain expertise to select different combinations of visual components and analyze the relationships between the different characteristics that contributed to the final data story and understand how the discoveries are derived from the data. I will demonstrate this in Section 6.2.

6.2 Case Study

In this section, I present a use case on crime report data. I then demonstrate how different analyses can be performed using different combinations of the visual components provided by the system to support decision-makers in law enforcement-related fields.

6.2.1 Data

I collected 7426 geo-tagged crime reports from Tippecanoe County, Indiana between June 1st, 2018 12:32AM and August 31st, 2018 11:17PM. In this use case, I highlighted three crime categories – trespassing, theft, and fraud for the topic-specific data distribution, the time-based analysis, and the correlation analysis visualizations. In addition, I also highlighted drug violation crimes to the time-based analysis. These topics were selected based on previous studies performed with the same dataset [52] and suggestions from domain experts.

In addition to the crime report data, the official US Census Data from 2010 is used to populate the census data choropleth map [2]. The landscape visualization uses map tiles from Openlayers [3]. Finally, the color schemes used in the system are selected from ColorBrewer [4].

6.2.2 Insights

Overview and original data distribution

Upon first glance at the system, as shown in Fig. 6.2, we can see the overall boundary of where the data is distributed. From the text around it, we can see that this dataset contains 7426 data entries, including 472 trespassing reports, 544 theft reports, and 305 fraud reports. Below the map view, we can also see in the textbox the time frame the data was collected and the percentage of the three crime reports in the dataset. The data distribution, visualized using the 5-class allports heatmap, also shows how the 7426 data points are distributed throughout the region and where most of the data points were reported from. This initial display, though not highlighting any analysis, demonstrates that by simply layering the overview and the data envelope, the system is able to provide users with a quick and comprehensive overview of the dataset. Without the annotated boundary, the heatmap

[2] https://www.census.gov/data.html
[3] https://openlayers.org/
[4] http://colorbrewer2.org/

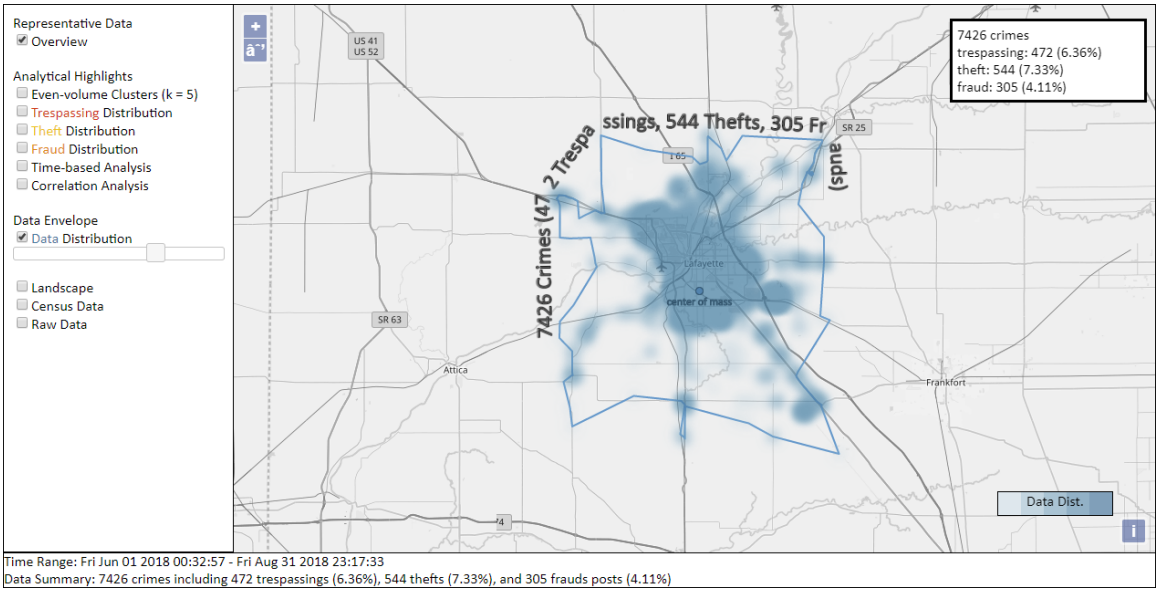


Fig. 6.2. An overview of a crime report dataset visualized with the annotated contour (Representative Data) and the 5-class allports heatmap (Data Envelope)

is not able to communicate the scale of the data effectively. Without the heatmap, users would have no way to examine the way the data is distributed throughout the region.

Even-volume clusters & data distribution

In Fig. 6.3, the system highlights the even-volume clusters with a k-value of 5. By adding the cluster boundaries to the overview and the data distribution, this separates the map into multiple regions, each containing an even volume of data entries and highlights the distribution hotspots within each region. This information can help a decision-maker in the police department better assess how to allocate resources and where to focus within each of the patrolling zones. The k-value can be adjusted based on the number of teams available.

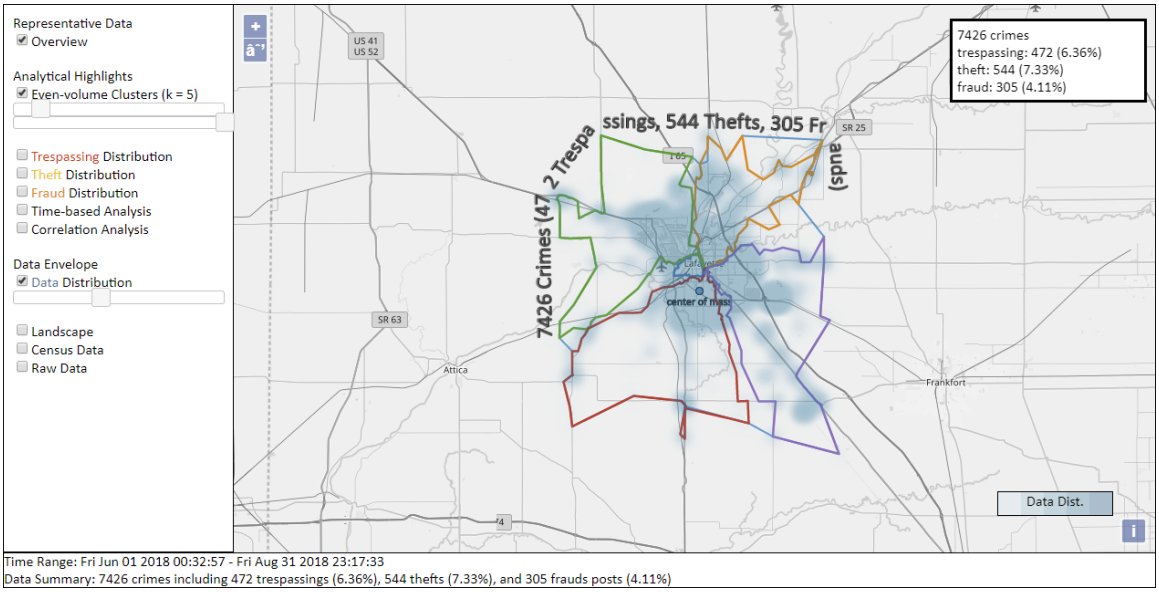
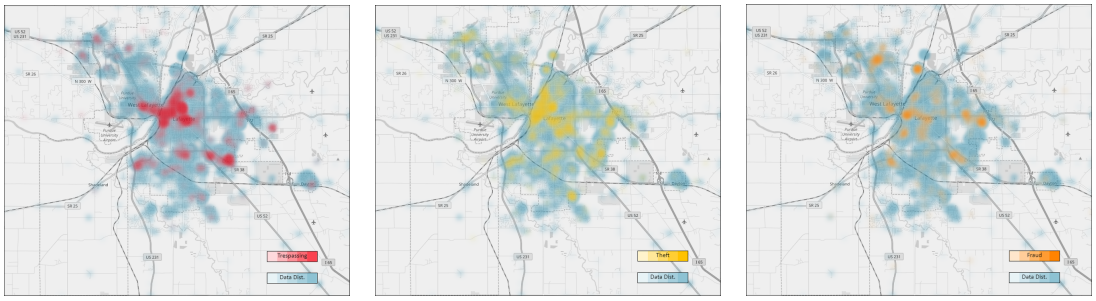


Fig. 6.3. Breaking the dataset into multiple regions using even-volume clusters (Analytical Highlights).



(a) Trespassing. (b) Theft. (c) Fraud.

Fig. 6.4. Crime distributions visualized using heatmaps (Analytical Highlights).

Crime distribution

Fig. 6.4 shows the distributions of the three highlighted crime categories. From the 5-class red heatmap in Fig. 6.4(a), we can see that trespassing crimes center mostly around downtown Lafayette and Purdue campus. From the 5-class amber heatmap in Fig. 6.4(b), we can see that theft crimes are more evenly distributed through the different neighbor-

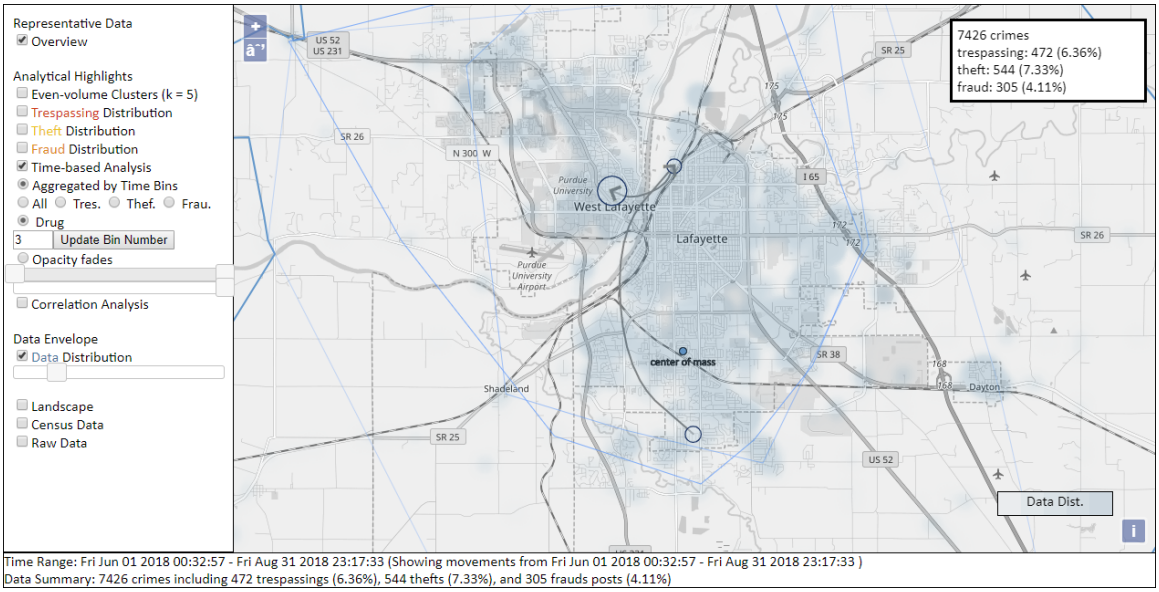


Fig. 6.5. Time-based analysis on drug violation reports over the summer of 2018 (Analytical Highlights).

hoods in the greater Lafayette area. From the 5-class flush orange heatmap in Fig. 6.4(c), we can see that fraud crimes focus on a few specific points. With the topic-specific distributions layered on top of the overall data distribution, we can see how each crime contributes to the overall dataset. If we add the cluster visualization back to the display, this can also aid the decision-maker in selecting teams with the right experience and training for the right neighborhoods.

Time-based analysis

Fig. 6.5 highlights how the center of mass and the count for drug violation reports moved from June 2018 to August 2018. Since the data was collected from the beginning of June to the end of August, by adjusting the bin number to 3, each circle on the map represents a month's worth of data. Based on the radii of the circles, it is clear that the number of drug violations increased significantly in the month of August. Based on the order and the positions of the circles, we can also see that the center of mass moved from the

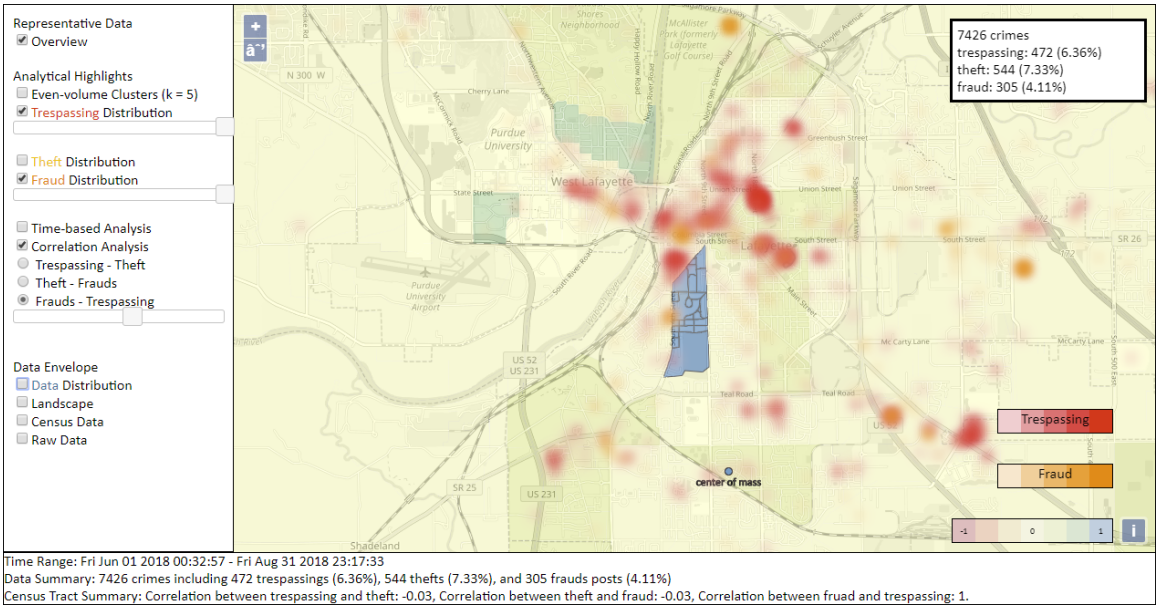


Fig. 6.6. Correlations between fraud and trespassing reports over the different regions are visualized using a 7-class spectral choropleth map (Analytical Highlights).

outskirts of Lafayette in June to the neighborhood besides Purdue University in August, the month when a new school year starts and students return to campus. With this information at hand, the decision-maker could further investigate the use of illegal drugs around specific neighborhoods and adjust the propaganda strategy against illegal drugs accordingly.

Correlation Analysis

Fig. 6.6 shows the correlations between crime reports on fraud and trespassing using a 7-class spectral choropleth map. In Fig. 6.6, we can see that fraud and trespassing have a strongly positive correlation in the south end of downtown Lafayette and a lightly positive correlation near Purdue campus. However, by overlaying the heatmaps of the two crime reports, we can see that the positive correlation is a result of a small portion of the two crime reports only. While both the distribution feature and the correlation feature are included in the system as an analytical highlights component, the different visualization techniques al-

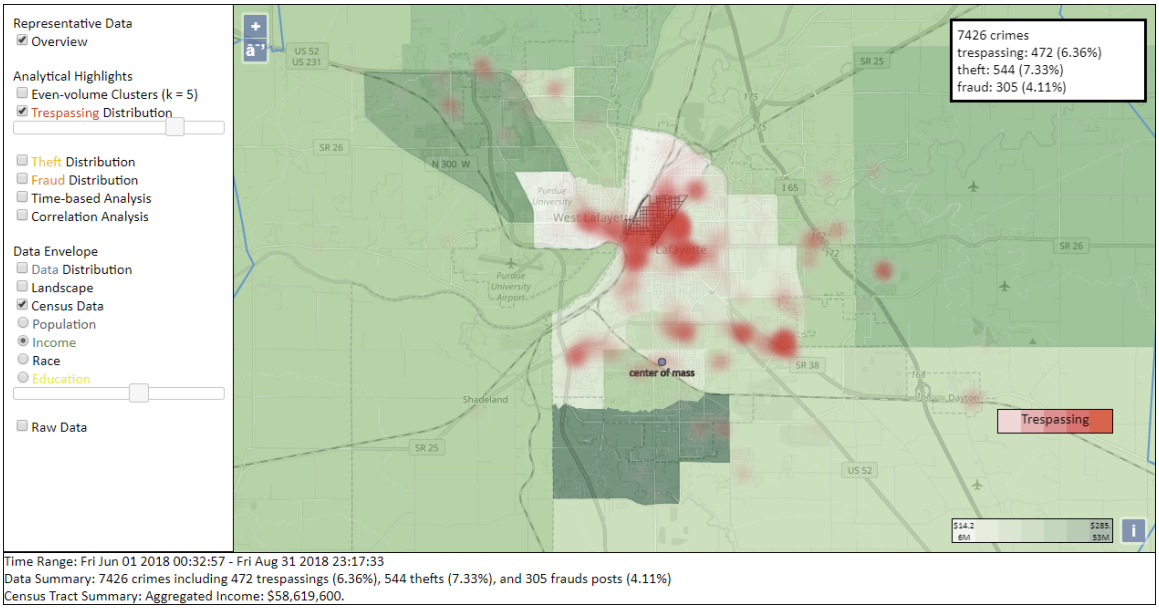


Fig. 6.7. Trespassing distributions (Analytical Highlights) on top of income data (Data Envelope) for additional context.

low the two features to be visualized simultaneously. In this case, the distribution provides additional context to the correlation results and allows users to reconsider the significance of the finding.

Census Data

Asides from the information encoded in the raw data, this design also imports additional context from the census data with the hope to explain certain phenomena or identify previously unknown correlations. Fig. 6.7 overlays the 5-class red heatmap for trespassing distributions on top of the 7-class green choropleth map for the aggregated income of each census tract. The figure suggests that trespassing primarily happened around neighborhoods with lower incomes. By updating the choropleth map to visualize the proportion of different ethnicities resided within each census tract, as shown in Fig. 6.8, we can see that there are three regions in Lafayette that have a slightly higher African American per-

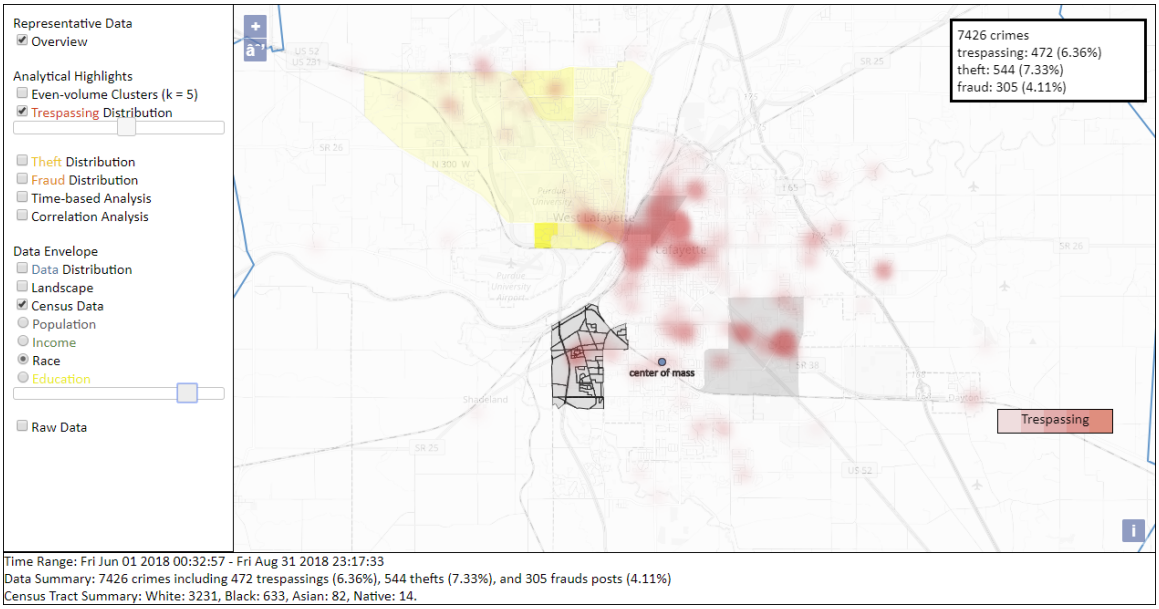


Fig. 6.8. Trespassing distribution (Analytical Highlights) on top of race data (Data Envelope) for additional context.

centage. The figure suggests that the triangle formed by the three regions attracted more trespassing crimes. It is worth noting that the insights gained through such visual examination deserve additional investigation before making any conclusions and that the correlation identified does not equate causation.

Landscape

Fig. 6.9 overlays the data distribution heatmap and the time-based data movement on top of a landscape map tile. This setup can help decision-makers understand how the landscape might be able to explain particular phenomena. In this specific use case, the landscape of Indiana seems to have no noticeable impact on either the distribution or the movement of the crime reports. This is likely because of the minimal terrain variations in the region.

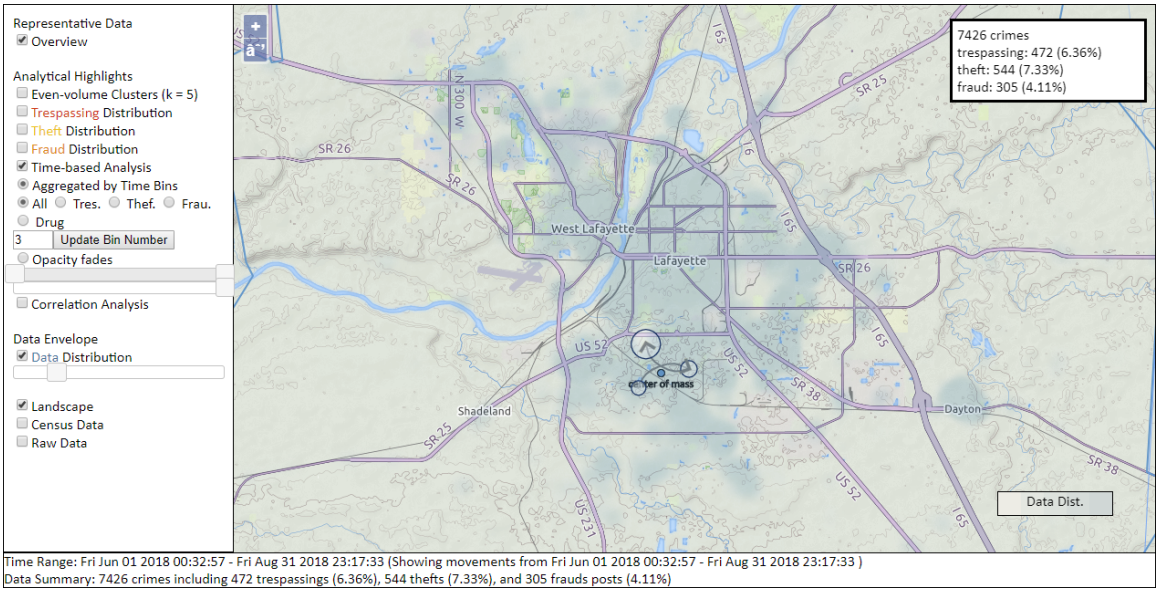


Fig. 6.9. Crime distribution (Data Envelope) and movement (Analytical Highlights) visualized on top of landscape map tile (Data Envelope) for additional context.

6.3 Feedback

This system was informally assessed by a high-level decision-maker in local law enforcement from Tippecanoe County, Indiana, in early February 2020. The decision-maker was familiar with the dataset and how the dataset has been utilized in a sophisticated visual analytics toolkit designed for crime analysts [44]. In this section, I summarize the feedback received after a demonstration of the system and an informal interview.

The decision-maker first explained that many decisions made in his department are "primarily incident-based" and mainly consider data relevant to specific criminal cases rather than the entire collection of reports, making this system not directly applicable to his daily practices. However, he found this system to be a valuable addition to the current practice, specifically in providing "a higher-level look at situational awareness" and "develop[ing] specific priorities within [the] common operating picture." The decision-maker considered the system a tool that can "help everybody [at the analysis and response center] understand

where problems exist and then develop a higher-level understanding of what the situation on the ground is, especially when it comes to the allocation of resources." He also recognized features such as the combination of even-volume clustering and data distribution to be a great tool for data-driven resource allocation, which, in his vision, the current practice will shift toward.

The decision-maker was able to confirm some of the patterns identified with his domain knowledge, such as the hotspot distribution of fraud reports tying to local businesses. He was also surprised by some of the insights discovered, such as the temporal analysis of drug violation, and was interested in investigating the different temporal patterns further.

The decision-maker especially appreciated that the system connected the data to the additional context of census reports and landscape. The external context was not well-utilized in the current practices of his department, and the decision-maker found the ability to understand the possible influences of cultural components to have great potential in allowing a better understanding of the data. The decision-maker also expressed interest in incorporating more context-focused map tiles into the system for further exploration, such as the correlation between streetlights and night-time crimes.

The main improvement the decision-maker wanted to see is the ability for the system to identify interesting patterns automatically. He explained that while the system can be effective for examining items he knows he is interested in, it is also important to be able to identify knowledge he does not know he needs to know. This request could be more beneficial to investigative fields similar to law enforcement, but is nonetheless a research problem worth further exploration.

Note that this summary is the result of an informal discussion with a domain expert on the usability and the significance of the system to his work. While written notes were taken during the discussion, no official report was generated. A questionnaire designed to mimic what users might be able to answer at the end of MILCs [158] can be found in Appendix B. While MILCs were not considered for this research work due to constraints in time and resources, this questionnaire was used to initiate the discussion.

6.4 Discussion

6.4.1 Strengths

The main advantage of this system is its simplicity. As a tool designed for decision-makers who are not professionally trained data analysts, it requires significantly less effort to use. It reduces possible confusion for its users, allowing them to obtain knowledge otherwise traditionally acquired through lengthy exploration. The system utilizes simple analyses that do not require expertise in data analysis to understand, opening the tool to a broader group of audiences. The audiences can perform every action through click, drag, and hover. The direct feedback from each of the checkboxes and sliders also makes using the tool more intuitive when setting up the display. With its simplified nature and design choices such as moving the lengthier text outside of the visual components, and allowing users to adjust the transparency of each layer and placing emphasis on different elements, the final display of the system can also effectively communicate insights from the data to other audiences.

Even though the system is simple, it still provides the basic analyses common to spatial data and covers the different factors that are important to the data story. Being able to use any combination of the analyses provided opens up more opportunities than focusing on one analysis at a time or going back and forth between multiple views. Many of the analysis results can be combined to provide additional contexts or insights, as shown in the case study.

The system utilizes both the users' domain expertise and the machine's computational power. It preserves the context during the simple exploration stage to allow users a stronger understanding of the reasoning behind the analysis results retrieved. The simplicity of the system and the pre-computation of the majority of the analyses provided also reduce the time required to retrieve knowledge compared to most tools. The addition to incorporate external context also allows a more comprehensive exploration of the dataset.

6.4.2 Trade Offs

The simplicity of the system, however, also leads to the inflexibility of the system. While the Three-Component Visual Summary design encourages customized design for its users, the users are limited to the functions and adjustable variables provided by the final system. As a result, the system may have limited contributions to professional data analysts and limits the potential for its users to grow into more sophisticated data analysis.

The web-based platform also limits the processing power of the system. This can be seen when the size of the dataset exceeds a certain threshold. The initial processing of the data becomes noticeably slower.

Overlaying components with area-based visual encoding can also be challenging. For example, a single context-specific hotspot works well with other visualization displays, but becomes harder to distinguish when multiple context-specific hotspots are displayed simultaneously. The adjustable transparency setting helps users to see visual components that overlap, but when heatmaps of different colors overlap, the transparency starts to mix the colors, which can cause additional confusion.

6.4.3 Scalability

Since the system does not examine the full content of the crime reports, the visual presentation (ignoring the processing power of a web browser) handles the scale of the data reasonably well. The contour visualization and the normalized heatmap can easily present a large amount of data, and the landscape is independent of the data. The main challenge in scalability is shown in the data movement visualization and the number of features utilizing visual designs of the same style being displayed at the same time. The first challenge is straight forward. As the user-selected number of temporal bins increases, the paths connecting the nodes start to overlap more and could become difficult to follow eventually. The second challenge is slightly more complicated. The visualization techniques used in this system can be roughly grouped into three different categories. The first category of visualization techniques overlays items directly on top of its location on the map. The

raw data points, the time-based analysis, and the landscape map tiles belong to this category. The second category of visualization techniques draws contours to surround an area. The overview and the even-volume clusters belong to this category. The third category of visualization techniques utilizes area overlay to demonstrate the change over space. The heatmaps for data distribution and the choropleth map for the correlation analysis and census data belong to this category. Techniques from different categories usually have little conflict with each other, and being able to adjust the transparency of each visual element made displaying multiple techniques from the same category, especially the last category, possible. However, when transparent layers with different colors overlay, the colors blend and could cause confusion. As a result, the display scales better when the selected analyses do not fall within the same category.

While the use case presented in this paper focuses on crime report data, the same approach and analyses can be performed over other geospatial data for useful insights. For example, social media data can be used to study and compare the topics of interest among different neighborhoods. This design of the system should be able to incorporate most geo-tagged multivariate data.

7. SUMMARIZING NETWORK DATA

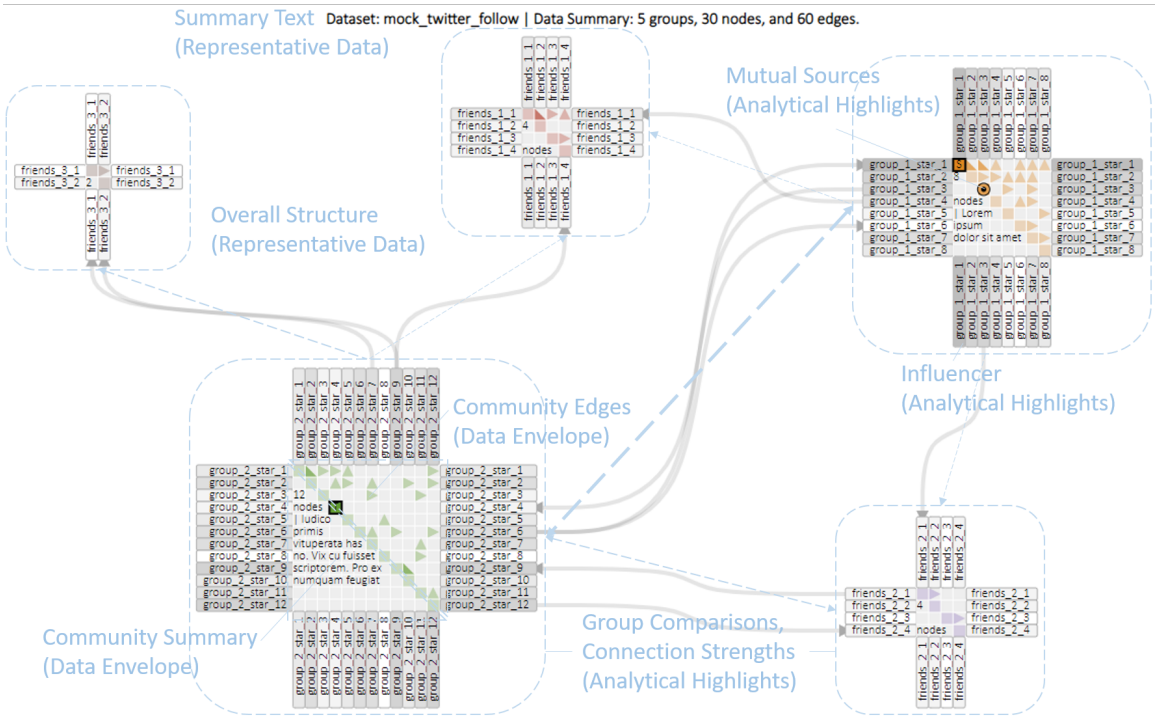


Fig. 7.1. A Three-Component Visual Summary design for a Twitter following network. The overall structure and the text summary serve as the representative data component, providing the audience with a quick grasp of the data size and the high-level network structure. The Analytical Highlights component covers network-relevant analyses including influencers, group dynamic comparison, connection strength and direction, neighbor distance, shortest path, and mutual data sources and targets. Network-level analyses are encoded into the design, while node-specific analyses require nodes to be selected. In this example, the mutual data source of the two selected nodes is being visualized. Finally, the edge information within each matrix and the community summary serve as the data envelope component to support the analyses with raw data and context.

With the popularization of the World Wide Web, social media introduced a new way for people to communicate. Users can generate and publish content easily using identifiable or anonymous accounts. Social Media content, some of which is unverified, can quickly spread and influence people across the globe, altering their opinions or behaviors. For example, many users now retrieve news stories from social media tools such as Facebook and Twitter instead of the traditional media outlet [159, 160]. Many also trade through online platforms such as eBay, Craigslist, and Facebook. With many of these interactions over the internet recorded and converted into digital data entries, more opportunities are provided to explore and analyze human connections and interactions. These opportunities can greatly benefit decision-makers in multiple domains. For example, understanding how information travels through different social media accounts can aid decision-makers in cybersecurity to identify possible sources of false information or dangerous propaganda. Knowing how news stories are shared between different organizations and individuals can help a decision-maker running a campaign to decide where to insert a piece of information to help it propagate quickly. Recognizing the direct and indirect connections between different identified criminals can help decision-makers in law enforcement identify potential key players behind the curtain and predict future targets.

By now, it is reasonably common to analyze network data using graph theory [102, 103]. However, most network visual analytics utilize the traditional node-link diagram [1], which can become visually cluttered quickly [111] and become challenging to navigate effectively if not familiar with the filtering and highlighting functionalities in interactive visual analytics systems. Many network visual analytics systems allow users to manipulate the layout of the node-link diagram and encode variables into different attributes of the diagram, such as colors, stroke width, etc. [2]. Utilizing this practice meaningfully will require an understanding of the effect the different variables have on the display, enough time to explore the different setup, and the training in analyzing a network using an interactive visual analytics system still. As we have learned from our domain experts, the decision-

[1] http://www.visualcomplexity.com/vc/index.cfm?domain=Social%20Networks

[2] https://video.sas.com/detail/video/6029136083001

maker survey, and Andrienko et al. [153], many decision-makers do not possess such time and training.

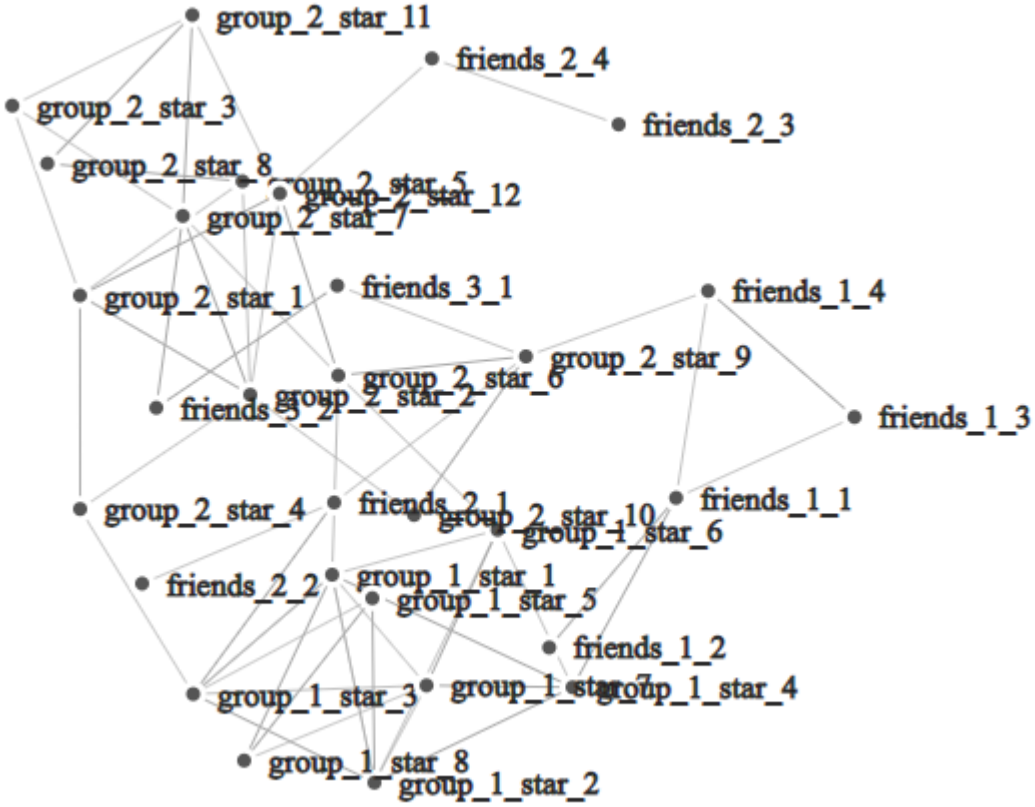


Fig. 7.2. A simple node-link diagram displaying the same network dataset shown in Fig. 7.1.

In this chapter, I present another web-based visual analytics system designed based on the Three-Component Visual Summary design to support casual-expert decision-makers in exploring and understanding data flow networks. I built my design on top of NodeTrix [14], which excels at providing the overview (representative data) while preserving the details (data envelope), by adding the missing analytical highlights component and introducing additional external context. This design utilizes the global structure of the NodeTrix layout and an additional summary text as the representative data, and the detailed connections displayed in the adjacency matrices and additional community descriptions as the data en-

velope. Added to this visualization design is a list of analytical highlights that are beneficial for network-related decisions on top of the NodeTrix layout, including influencer, neighbor distance, shortest path, mutual sources and targets, connection strength and directions, and group dynamics comparison. Node-specific analytical highlights can be toggled on and off using click events. Hover events are also available to help users examine details or highlight a selection. Shown in Fig. 7.1 is the design visualizing a mock-up Twitter Following network that contains 5 communities, 30 nodes, and 60 edges and highlighting the influencers and the Twitter account the two selected Twitter accounts both follow. In comparison, Fig. 7.2 displays the same dataset in a traditional node-link diagram. The difference in visual clutter is clear.

In the following sections, I explain the design choices of the three components and the system, present interesting findings through a case study and the feedback from decision-makers that are domain experts of network data, and discuss the outcome of the design.

7.1 Design

In this section, I explain the design of my Three-Component Visual Summary for network data. This design focuses on visualizing data flow networks to support decision-makers in understanding how information spreads and identifying influencers (sources that are sending data to a large number of receivers directly) and relationships between different actors or communities. When data flows from node A to node B, it is visualized by an edge or a glyph pointing from node A to node B. In the example presented in Fig. 7.1, if Twitter account C follows Twitter account D, the data is flowing from node D to node C and will result in a visualization where node D points to node C.

7.1.1 NodeTrix

Henry et al. presented NodeTrix in 2007 and introduced the first network visualization design that combines the strengths of the node-link diagram and the adjacency matrix representation [14]. NodeTrix utilizes the two visual representations of network data to show

the high-level global structure of a network and support exploring the low-level connections within the communities effectively. This design direction parallels with the representative data and the data envelope component of my Three-Component Visual Summary design. The next logical step is to highlight analysis results in the visual display, instead of obtaining them through an extensive interactive exploration process. Therefore, I decided to build my network visual summary design on top of the NodeTrix design.

One major difference between my application and NodeTrix is in the dataset it targets. While both applications handle network data with communities, NodeTrix visualizes weighted non-directional data and manipulable communities such as co-authorship between different research groups, and my work focuses on non-weighted directional data such as Twitter following among different ground-truth communities. I decided to group the nodes based on real-life communities to preserve context rather than reducing crossing edges with automatic clustering algorithms. As a result, the authoring support in NodeTrix that allows users to understand how community matrices are formed is not carried over to this iteration of my network visual summary design. Given datasets with multiple node-based attributes, however, animated transformations between different user-selected ground-truth communities may be implementable to achieve a similar effect. I leave this to future work.

7.1.2 Representative Data

The representative data component in this design combines a global node-link diagram structure from the NodeTrix layout and a short summary text. The summary text is placed on the top of the design and displays the name of the dataset and the number of communities, nodes, and edges. Unlike the proposed geospatial design, the summary text is not necessary for users to extract the number of nodes and edges. However, the summary text allows users to understand the scope of the dataset more quickly, satisfying the design requirement of the Three-Component Visual Summary design. The global structure of the network can be seen through the node-link diagram, where each node represents a commu-

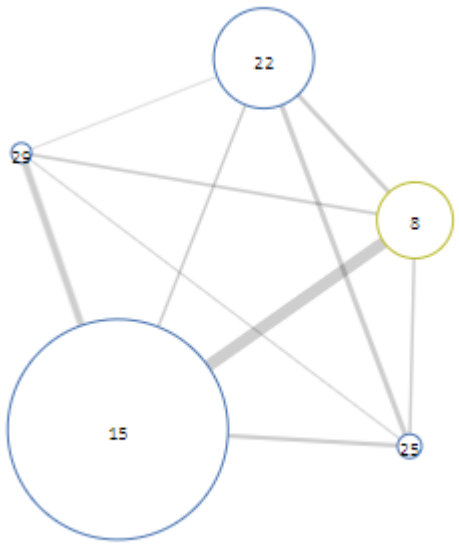


Fig. 7.3. The system also includes a zoom-out level of the visualization to accommodate large scale datasets. This figure visualizes an e-mail network between members from five different departments of a research institute. The dataset is further explored in the case study.

nity. First, users can obtain an easy grasp of the scale of the nodes and if the nodes connect to each other. From each node in the diagram, visualized as an adjacency matrix, users can roughly understand the size of the communities. From the number and the directions of edges between the nodes, users can also see the relative strength and direction between the communities.

A simple node-link diagram is also drawn as a ”zoom-out” overview display, as shown in Fig. 7.3. Each node represents a community. The radius of the nodes reflects the size of the communities, the width of the edges represents the combined non-directional connection strength between the communities, and the texts in the center of the nodes are the community ids. This is designed to accommodate datasets that include a large number of communities and cross-community edges. The node-link diagram was chosen for the zoom-out overview instead of an adjacency matrix representation because of the audiences' familiarity [14] and the resemblance to the zoom-in visual.

7.1.3 Analytical Highlights

The analytical highlights component provides a set of analyses that help users further explore and understand the key players and the relationships between different nodes and communities in a network dataset. The network-level highlights are incorporated into the overall design. The node-specific highlights are triggered by selecting a node or a pair of nodes through clicking events. No query is needed to select the nodes.

The network-level highlights visualize the analysis results that are constant regardless of the node being examined. This includes the influencers, the community dynamic comparison, and the connection strengths between communities. Being able to identify the influencers of a network can be crucial for maximizing the speed to diffuse a piece of information [161]. In the scope of this work, how influential a node is is defined by the number of nodes directly receiving data from it. The influence score for each node is normalized across the dataset and encoded into the label color of the nodes in the adjacency matrices to allow users to identify the most influential nodes in each community and across the dataset.

In the NodeTrix layout, both the label of a node and its corresponding cell in the diagonal line of the matrix (where both the row and the column represent the node) can be used to encode the influence score. The system encodes the normalized influence score using the opacity of the label color to reduce the effort in identifying the node and to match the design of NodeTrix – the darker a label is, the more influence the node has. Graph comparison is another form of analysis that can impact a wide range of domains [162]. With each community containing a sub-network, comparing the dynamics between different communities can generate useful insights. While comparing node-link diagrams visually can be challenging, high-level visual comparisons between different adjacency matrices are more straightforward – how do the different nodes in each community communicate with each other, does a community mainly send out information, receive information, or both, etc. As a side product of the NodeTrix layout, the connection strength between two communities is visualized through the combination of the count and the directions of the edges connecting the two communities. All of the edges connecting two communities are designed to connect to and from the same side of the corresponding matrices, using the same two-turn curve to help users separate the edges connecting different communities. Each edge is semi-transparent in order to allow users to identify overlapping. The connection strength gives users insights into the relationships between the different communities. It is easy to identify the lack of connection or minimal connection between communities. A strong connection represented by a narrower but darker-edged overview means the communication mainly goes through specific members, whereas a strong connection represented by a wider but lighter-edged overview means the communication happens more on personal levels. By correlating the understanding of the comparisons and the connections between different communities to additional context, decision-makers from management can identify patterns that lead to more successful outcomes.

The node-specific highlights can be triggered when users select either a node or a pair of nodes. These highlights focus on the relationships of specific nodes. When a node is selected through clicking its corresponding cell in the matrix, cells that represent all the nodes it can, directly and indirectly, send information to will be marked with the neighbor

distance. This allows decision-makers to further explore the diffusion of information [161] – how much direct influence does the player have, how far can information reach, and what are the distance or time it takes for the information to travel to the targeted receiver. Distance-1 and distance-2 neighbors are highlighted with a higher opacity, neighbors with a distance of 3 and beyond have a lower opacity, and nodes that are not reachable are dimmed out. By visualizing the neighbor distances over the matrix cells, users are able to select the next node of interest by clicking directly on the component that presents the information which motivated the action. This design reduces the effort to navigate through the matrix representation for node selection. Additionally, edges that do not connect to the selected node will also be dimmed out to focus on the cross-community influence of the node. Hovering over the neighbor distance of a node will trigger the system to display the shortest path in a tooltip. When two nodes are selected, the system highlights the mutual distance-1 sources and targets of the two nodes using arrowhead and arrow feather glyphs that are commonly used in Physics to represent vector into page and vector out of page. This supports decision-makers in detecting second-degree contacts between two seemingly unrelated players, which can be important to domains such as crime network investigation [163]. The mutual source uses the arrowhead glyph as the data comes out of the node, and the mutual target uses the arrow feather glyph as the data goes into the node from the two selected nodes. This visual representation is chosen for three reasons. First, the glyph needs to be easily identifiable in the small area defined by the cell. Second, the glyphs need to be able to stack on top of each other for nodes that are both the mutual source and the mutual target of the selected nodes and remain recognizable as the combination of the two glyphs. Finally, while instructions are provided, it is ideal to select a design that is used in different fields that users might recognize to help them make the mental connection easier. An alternative design uses letters to encode mutual sources and targets (e.g., 'ms' and 'mt'). However, a larger grid size will be required to encode recognizable letters. Additionally, the amount of space required for the complete label to eliminate the need of instruction is impractical. With instructions required for both options, the glyphs were

chosen to reduce the space necessary. Hovering over the target node will again display the shortest path from the source.

7.1.4 Data Envelope

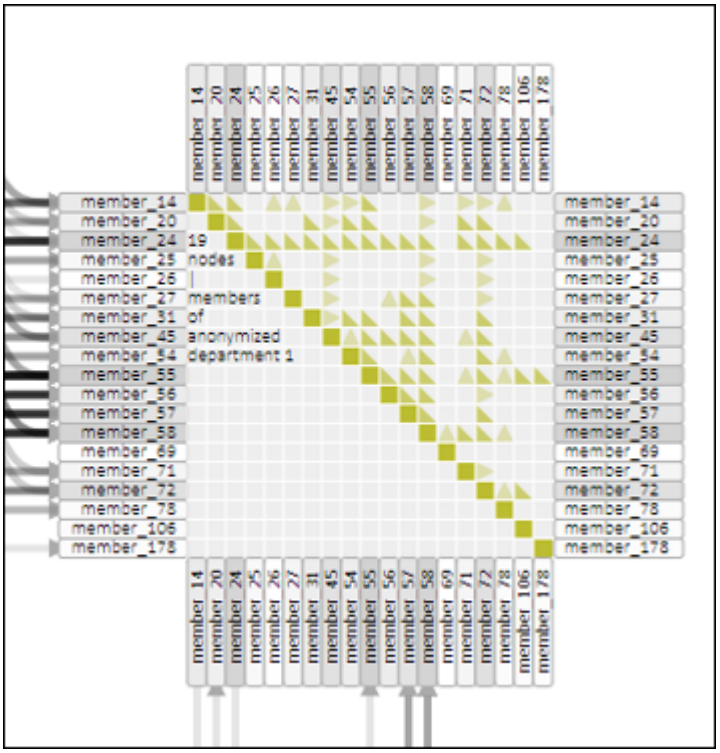


Fig. 7.4. Three directional glyphs are added to the matrix cells. One glyph points from the row label to the column label, one glyph points from the column label to the row label, and one glyph connects both ways. The lower half of the matrix provides a summary and a description of the community.

The data envelope component allows users to explore the dataset further and compile reasoning to the analytical highlights. Having the ability to understand the context of each community and trace the connections within can help users validate the insights gained and explore potential reasoning behind the findings. With the design built on the NodeTrix layout, this is mostly achieved through the updated design of the community matrix, as shown in Fig. 7.4.

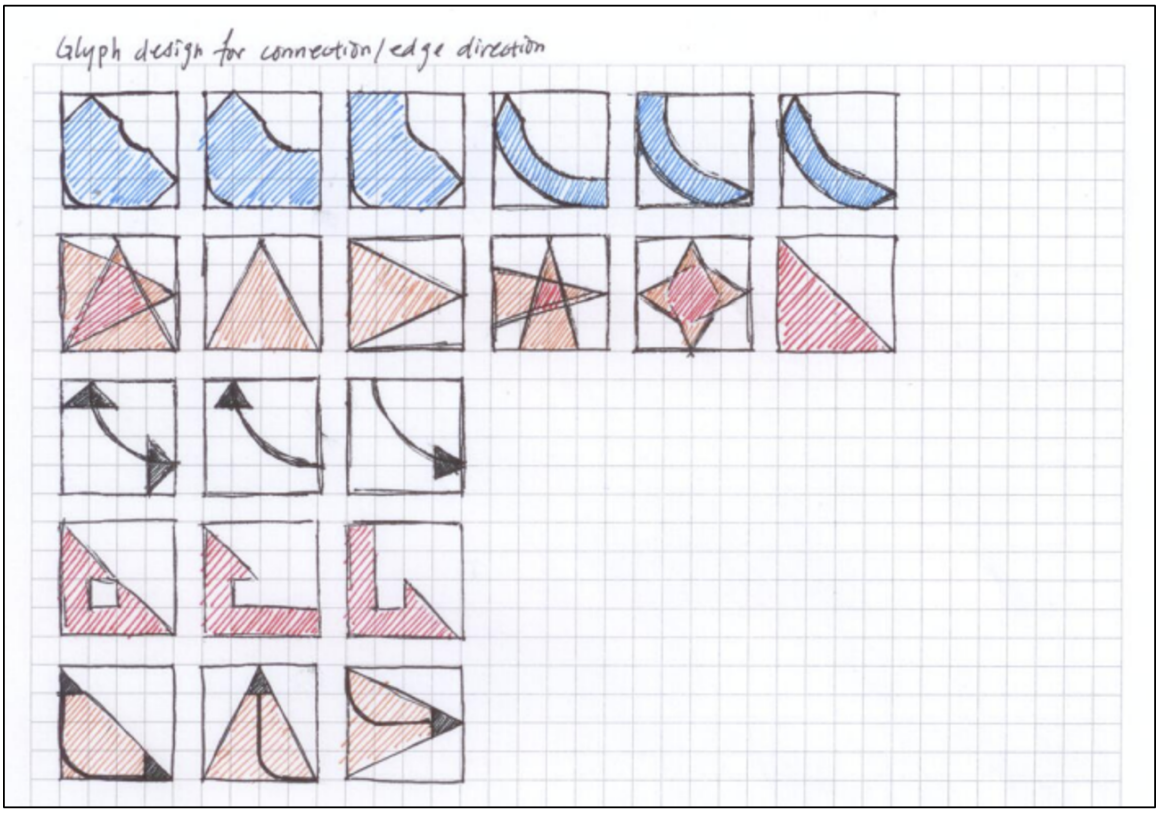


Fig. 7.5. Different glyph designs were considered to indicate the direction of the information flow within the community adjacency matrices.

As described earlier, this design focuses on non-weighted directional edges. While adjacency matrix can encode direction using the row-to-column direction, this may not be known or intuitive to the casual experts. Therefore, I keep the matrix symmetric (like the NodeTrix design that focuses on non-directional edges) but add directional glyphs to the cells inside the matrix to visualize the direction between the connected nodes. This allows users to visually and more intuitively trace if two nodes from the same community are connected, and if so, with which direction. Fig. 7.5 shows the different glyph designs considered to encode the connection direction in the upper half of the adjacency matrix. There are a total of three directions that need to be encoded: information flowing from the column node to the row node, information flowing from the row node to the column node,

or information going both directions. An important factor to consider is that the glyph must remain easily recognizable when the grid size of the matrices is small. As a result, glyph designs from row two column two (row-to-column), three (column-to-row), and six (both directions) were selected for their simplicity and well use of space while being discernible. From our initial feedback, the small sample of prospective users was able to utilize the selected glyph designs effectively after simple instructions. While the glyph designs are easy to learn, future studies on the intuitiveness and the effectiveness of different designs may be beneficial.

Since the community matrix design is kept symmetric, and the directions of the edges are encoded with glyphs to reduce the learning curve, half of the matrix is used to repeat the same information. To use the space more effectively, this design clears up half of the matrix to display a text description of the community to provide additional context to users. Similar to the geospatial application, additional information about the community can be added to the description text as the external context. However, the size of the matrix varies based on the number of players in the community, meaning matrices that represent smaller communities may not have enough space for the full text. To address this issue, I add a tooltip that will display the full text when hovering over a community description. It is worth noting that it is easier to examine the individual nodes in a smaller community, making the description for that community less critical to the understanding of the community.

7.1.5 System

Similar to the previous designs, the Three-Component Visual Summary design for network data also takes the form of a light web-based visual analytics system that allows the use of constrained interaction to manipulate and generate static outputs that satisfy the three-component design. The system can be populated using spreadsheets that store network data of interest in a structure commonly used for network visual analytics datasets. A set of simple instructions is provided on the top of the web page. Aside from the instruc-

tions, the system utilizes just one view, allowing users to layer and interact with the visual components directly and creating a more intuitive user experience with the direct manipulation feedback. Since the NodeTrix design largely covered the component designs of representative data and the data envelope, this work focuses on incorporating the additional analysis results into the visual presentation.

The system also provides the following interactive functionalities. Users are able to adjust between the two zoom levels. The zoom-out level presents a community-based simple node-link diagram, as shown in Fig. 7.3. The zoom-in level presents the three-component design described in the previous sections, as shown in Fig. 7.1 and Fig. 7.6. The initial positions of the matrices are determined using a force-based layout [124], but users can click and drag to move the community matrices around as they see fit, and the edge positions will update accordingly. Hovering the cursor over any cell in the matrices will highlight the labels of the corresponding nodes. Hovering over a path will highlight the edge with wider stroke width and a darker arrow, and trigger a tooltip to present the source node and the target node. Hovering over the community description text in a matrix will trigger a tooltip to display the full text. Clicking on a diagonal cell in an adjacency matrix will select or deselect the corresponding node. At most, two nodes can be selected at the same time. If two nodes are already selected, the first selected node will be deselected when a third node is selected. When two nodes are selected, the mutual distance-1 source and target nodes are marked using arrowhead and arrow feather glyphs. Common signifiers are used throughout the design, such as cursor icons for panning, clicking, and dragging.

7.2 Case Study: E-mail Network

In this section, I present a case study using the Three-Component Visual Summary design for network data. First, I introduce the dataset that is being visualized. I then demonstrate the insights gained through using this visualization and how such insights can support decision-makers in performing data-driven investigations.

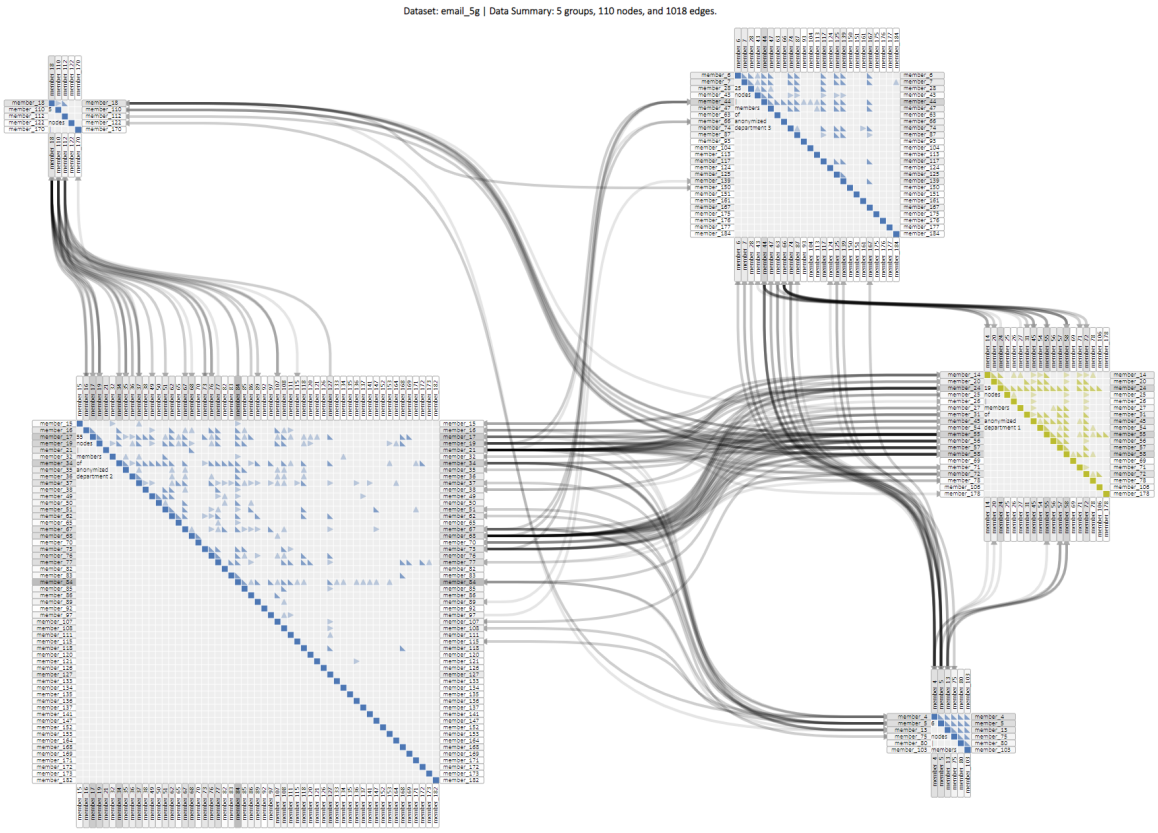


Fig. 7.6. An e-mail network visualized using the Three-Component Visual Summary design.

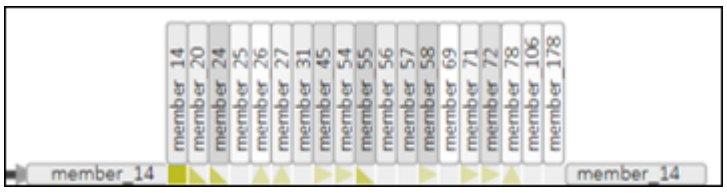
7.2.1 Data

Shown in Fig. 7.6 is an e-mail network between members of five departments in a European research institution over a period of eighteen months [164]. This is a self-contained, directed, non-weighted dataset with ground-truth communities. E-mails sent to or received from addresses outside of the five departments are excluded. The direction of the edge goes from the sender to the receiver. I am only interested in whether a connection has been established and do not include the count of e-mail exchanges between two e-mail addresses. Finally, the e-mail addresses are grouped by the department their owners belong to. This dataset includes 5 departments, 110 e-mail addresses, and 1018 established connections. The e-mail addresses and the department names are anonymized. The nodes of this network represent the owners of the e-mail addresses, the edges represent the existing history of at least an e-mail exchange between two e-mail addresses, and the communities represent the different departments. This differs from the Twitter following social media network example shown in Fig. 7.1, where when account A follows account B, the data flow direction is reversed to point from node B to node A, if e-mail address C has ever sent an e-mail to e-mail address D, the data flow direction points from node C to node D.

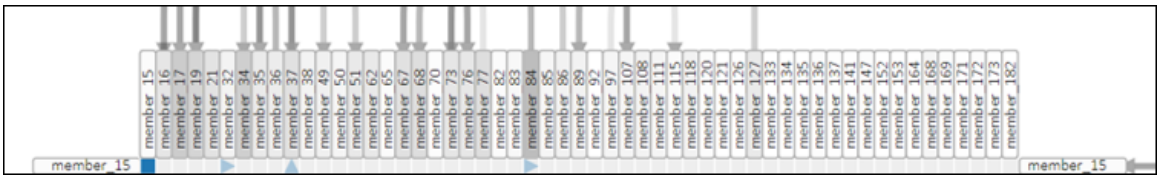
7.2.2 Insights

Influencers

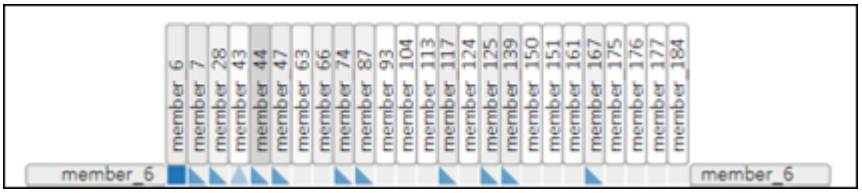
In Fig. 7.7, we can see the labels of each node in this dataset. Since the darker a label is, the more influence its node has, we can tell that member 84 from department 2 has the most influence across this dataset as well as department 2 (Fig. 7.7(b)). Examining the remaining departments independently, we can see that member 24, 55, and 58 are the more influential members in department 1 (Fig. 7.7(a)), member 44 is the more influential member in department 3 (Fig. 7.7(c)), member 4 and 5 are the more influential members in department 4 (Fig. 7.7(d)), and member 18 is the more influential member in department 5 (Fig. 7.7(e)). Knowing this information could help decision-makers understand the power



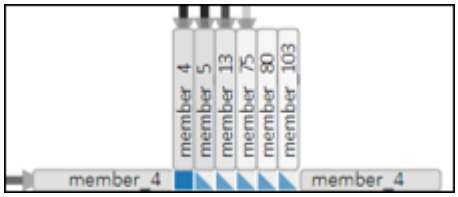
(a) Members of department 1.



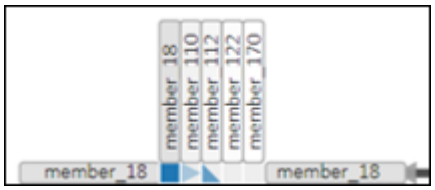
(b) Members of department 2.



(c) Members of department 3.



(d) Members of department 4.



(e) Members of department 5.

Fig. 7.7. The color of the labels help identifying the influencers of the dataset and of the communities.

structure of each department and identify members to reach out to speed up the diffusion of information.

Neighbor distance and shortest path

When a node is selected, neighbor distance is added to the display, and the shortest path to a reachable node is enabled through the use of a tooltip. If we select the node for member 84, the most influential node of the dataset, we can see that there exist members

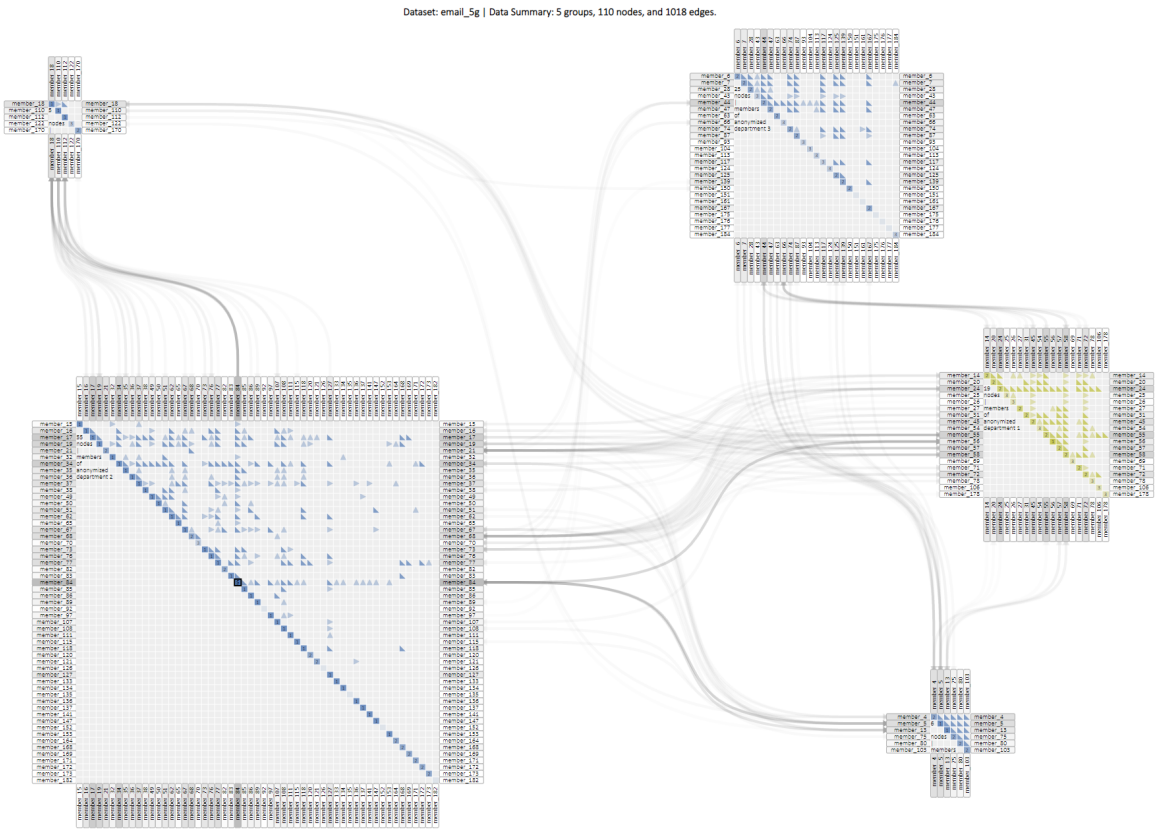


Fig. 7.8. Neighbors of node 84.

in every department besides department 3 that receive first-hand information from member 84. (Fig. 7.8). In comparison, member 170 has not sent an e-mail to any members of the five departments in the eighteen months (Fig. 7.9). Examining another member, member 122, in the department 5 member 170 is from, we can see that while information can flow from member 122 to the rest of the members in the department, none of them receive from member 122 directly (Fig. 7.10). In order for member 122 to spread information to the group, the information has to go through members in different apartments before circling back around. We can also see in order for member 170 to receive information from member 122, the fastest way based on the current communication network is to go through member 4 and member 37. With these insights, decision-makers can understand personal circles and how that may or may not be affected by the community a member is

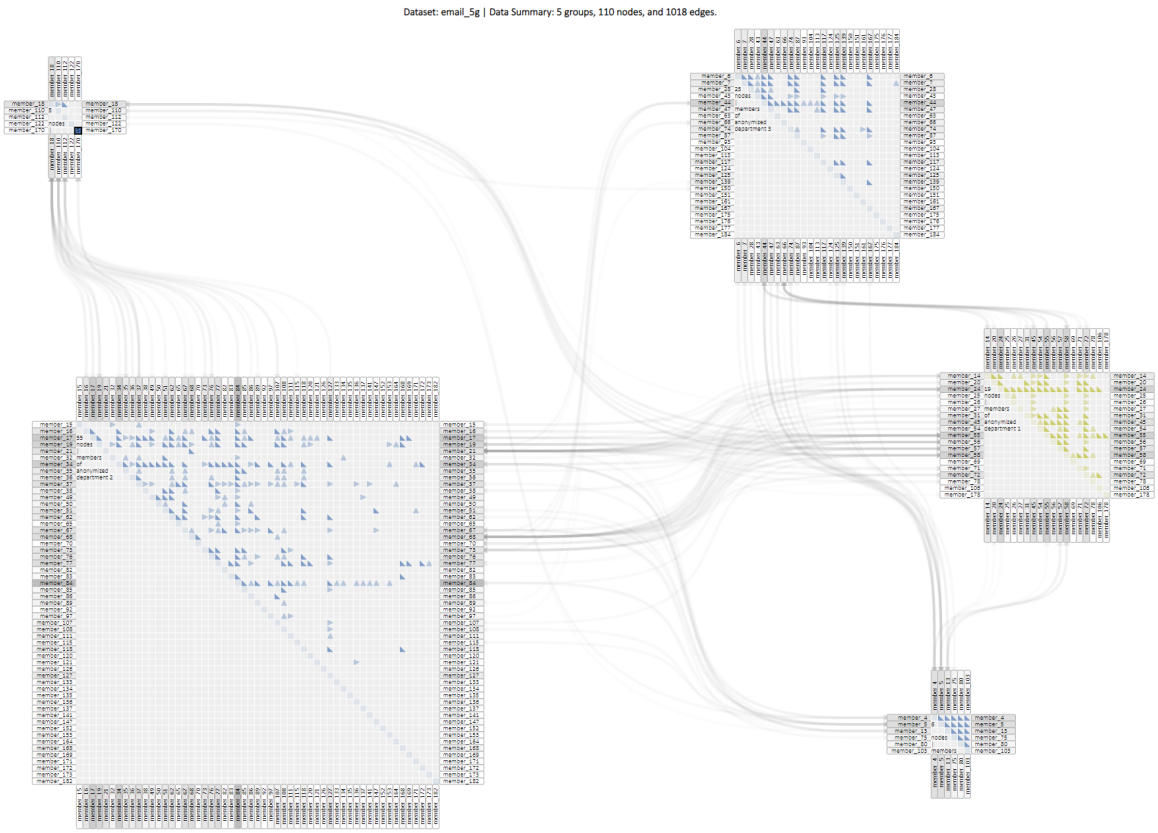


Fig. 7.9. Neighbors of node 170.

in. Additionally, decision-makers can use this information to filter out members that do not have connections to a specific member of interest. Finally, by examining the largest neighbor distance, decision-makers can estimate how efficient it is to communicate within the communities.

Mutual data sources and targets

By exploring the mutual sources and targets of two nodes, decision-makers can further understand the connections between the two nodes. In Fig. 7.11, we can see that after selecting member 14 and member 25 from department 1, their mutual sources and targets are highlighted. The two members are from the same department with a neighbor distance

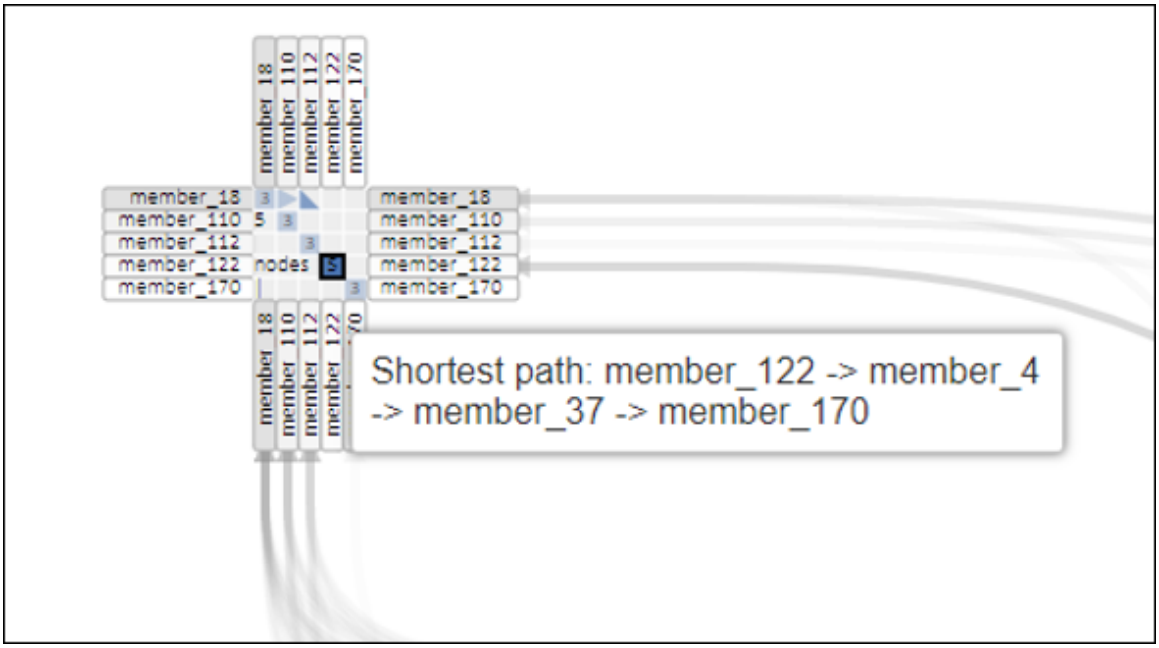


Fig. 7.10. Neighbors of node 122 and shortest path to node 170.

of 2 to each other. The two members both receive information from member 24, 45, 58, and 72, and they both pass information to member 24 and 26. Assuming member 25 and member 54 have shown similar noteworthy behavior, this may allow decision-makers to explore possible causes and take actions accordingly. Now, we explore the connections between member 6 from department 3 and member 14 from department 1. The two nodes are distance-2 neighbors from two different departments with no direct connection. In Fig. 7.12, we can see that member 44 is the only node that connects the two members, and member 6 and member 14 both send and receive e-mails from member 44. This indicates the possibility that member 6 and member 14 could communicate through member 44 even if they are not directly connected in the data flow network.

Applying similar analysis to a criminal investigation use case, decision-makers could use such insights to determine if two seemly unrelated players took similar actions because they are both under the influence of another character? Is there someone in the middle

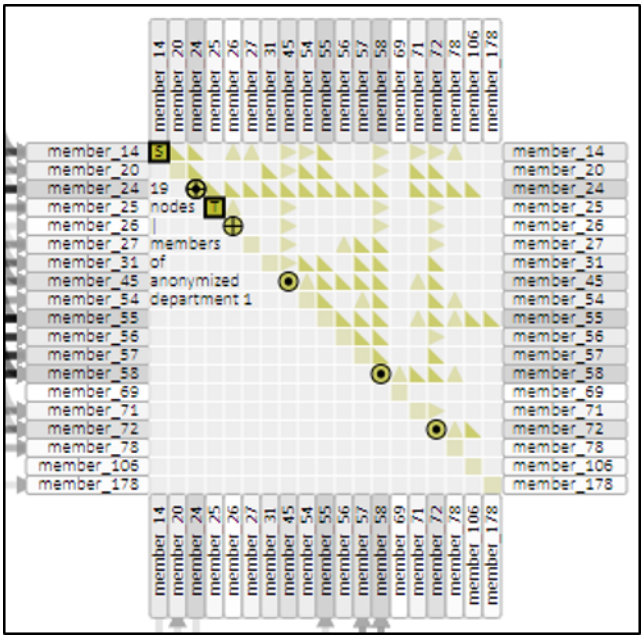


Fig. 7.11. Mutual sources (sender) and mutual targets (receiver) are highlighted when two nodes are selected.

that they are communicating through? If they are working together, who might be the next target?

Community connection strength and direction

By comparing Fig. 7.13(a) and Fig. 7.13(b), we can tell department 1 and department 2 have a strong and two-way relationship while department 3 and department 5 have a weak one-way connection. Due to the data anonymization, we do not know specifically what departments these are, which makes it impossible to use the context to explain the observation. We do know that it is common for different departments to collaborate on delivering products, and it is important to have good communication between the departments (e.g., Designers and Engineers). Understanding the connection strength and direction between the collaborating communities can help decision-makers explore whether an issue in a project could result from miscommunication between the different departments.

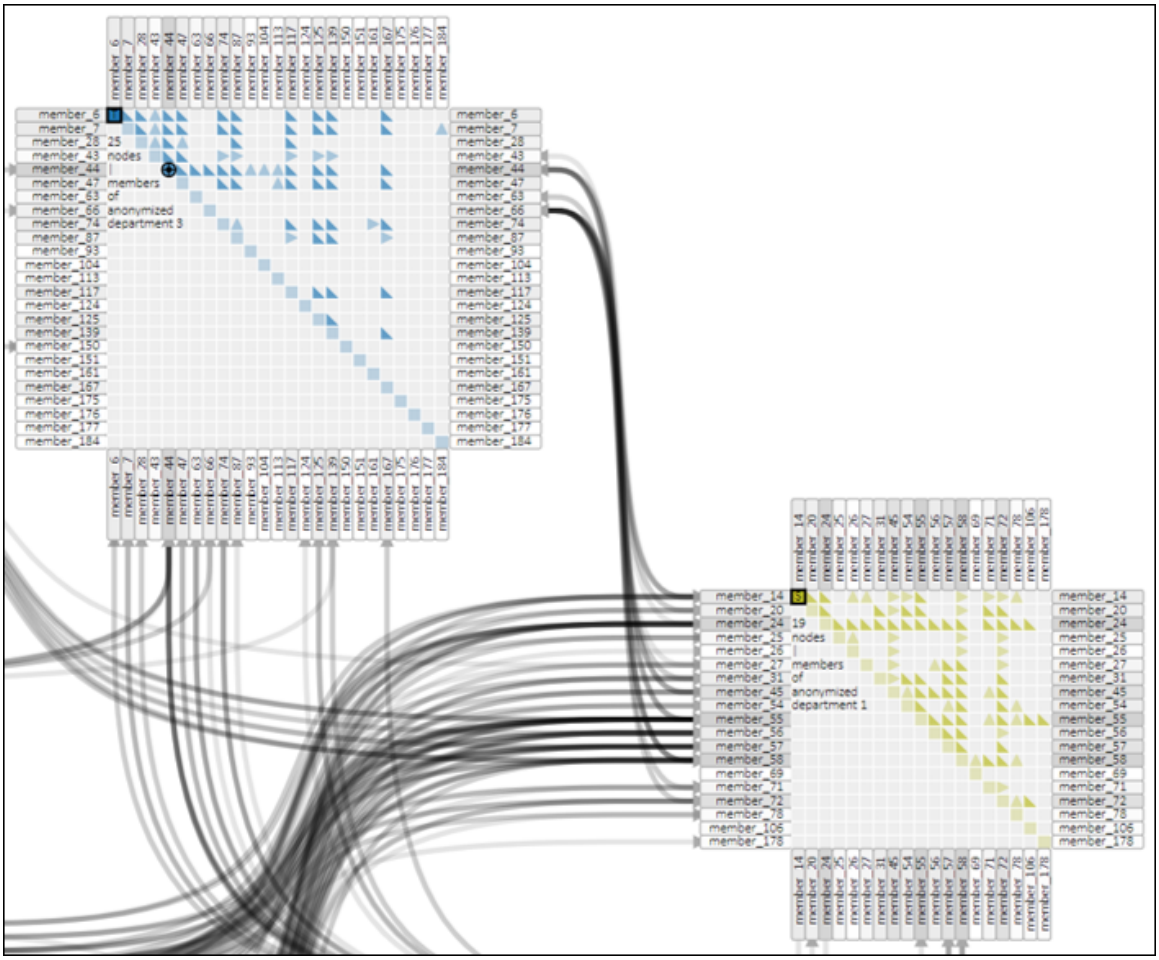
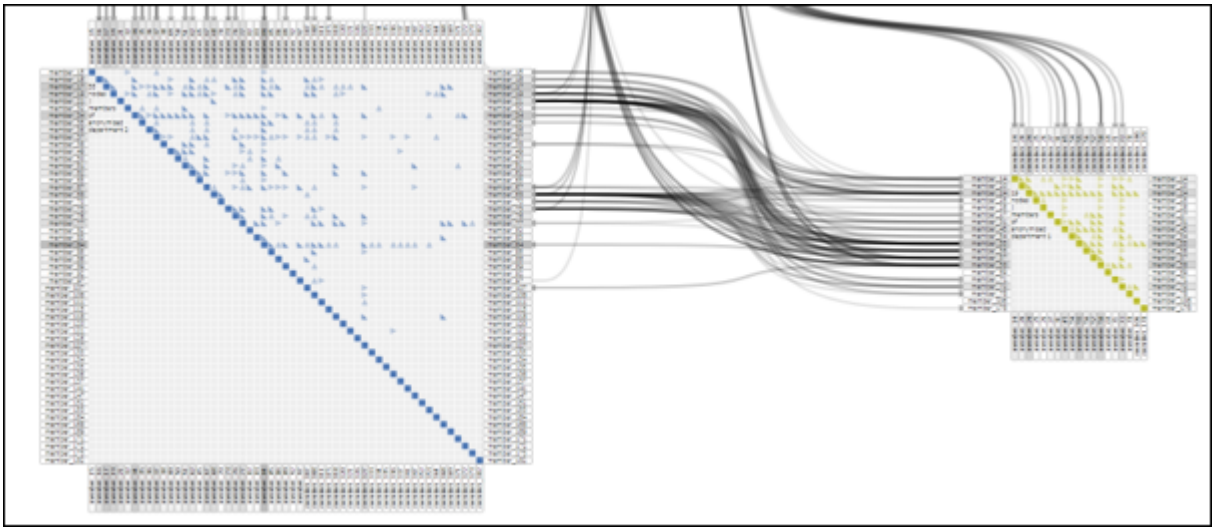


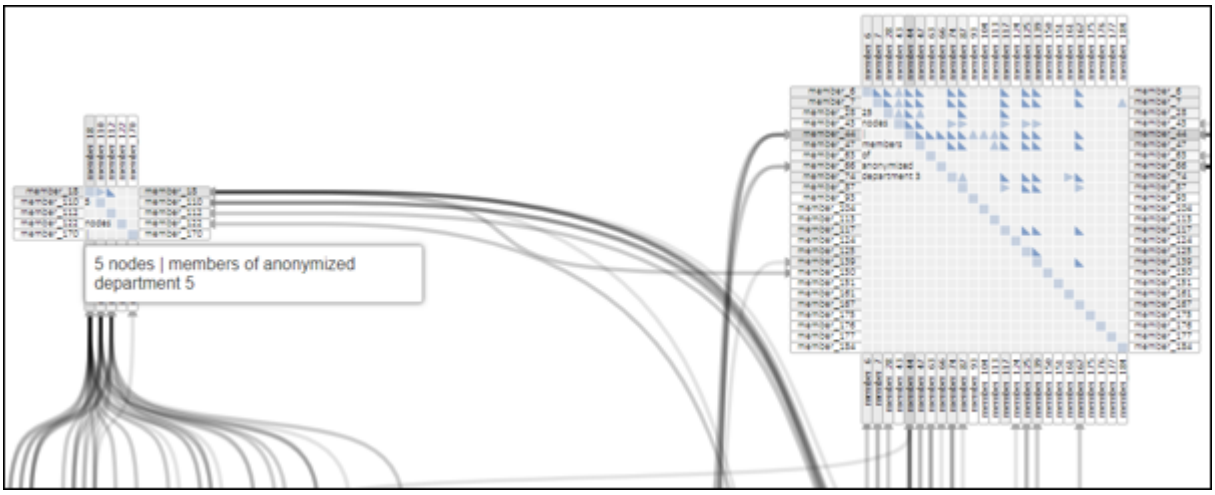
Fig. 7.12. Member 6 and member 14 from different departments have one mutual source/target.

Community dynamic comparison

In Fig. 7.14, we compare the group dynamics of the two smaller communities in the dataset. In Fig. 7.14(a), we can see every member talks to every member in department 4. On the other hand, Fig. 7.14(b) shows that there is very little inside communication between the members of department 5. By comparing the group dynamics with additional performance metrics, this could help decision-makers evaluate and identify patterns that lead to



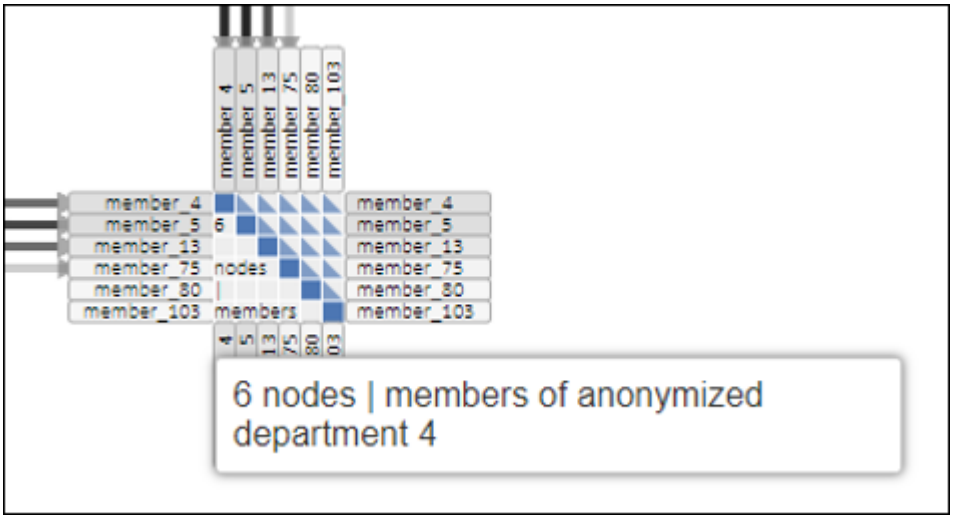
(a) Connections between department 1 and department 2.



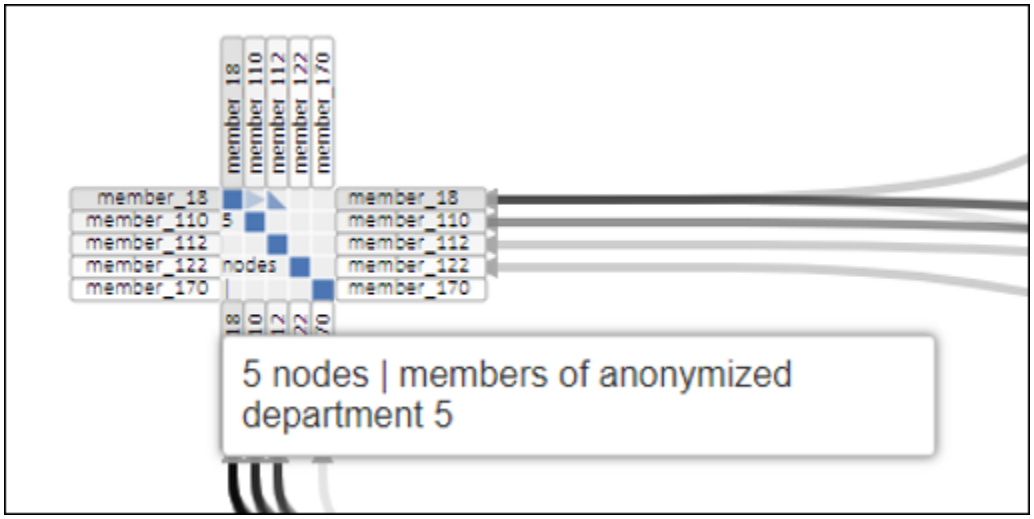
(b) Connections between department 3 and department 5.

Fig. 7.13. Connection strength and direction can be determined by examining the combination of edges between two communities.

a more productive work environment. If available, additional context such as performance metrics could be added to the community description to support this process.



(a) The group dynamics of department 4.



(b) The group dynamics of department 5.

Fig. 7.14. The dynamics between different communities can be compared by examining the differences in their matrices.

7.3 Case Study: Physician Network

In this section, I present a second case study using the Three-Component Visual Summary design for network data. This dataset contains additional metadata on the nodes that allow different community groupings. Through this case study, I demonstrate the effect of

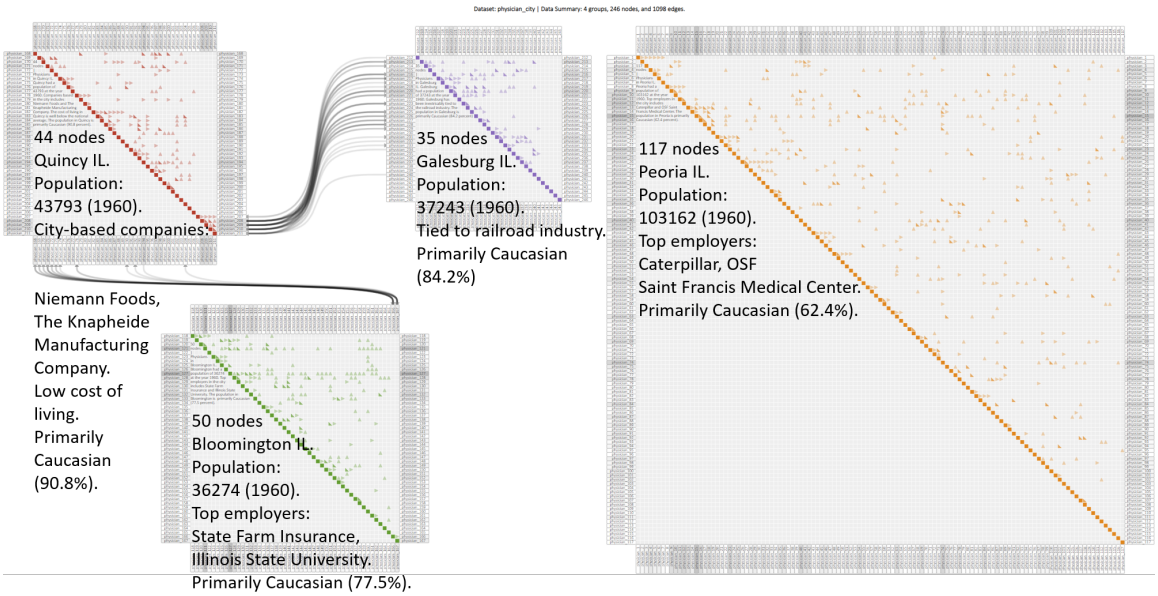


Fig. 7.15. A physician network, grouped by practicing city, visualized using the Three-Component Visual Summary design.

different community selections and external context, which was not available in the previous case study.

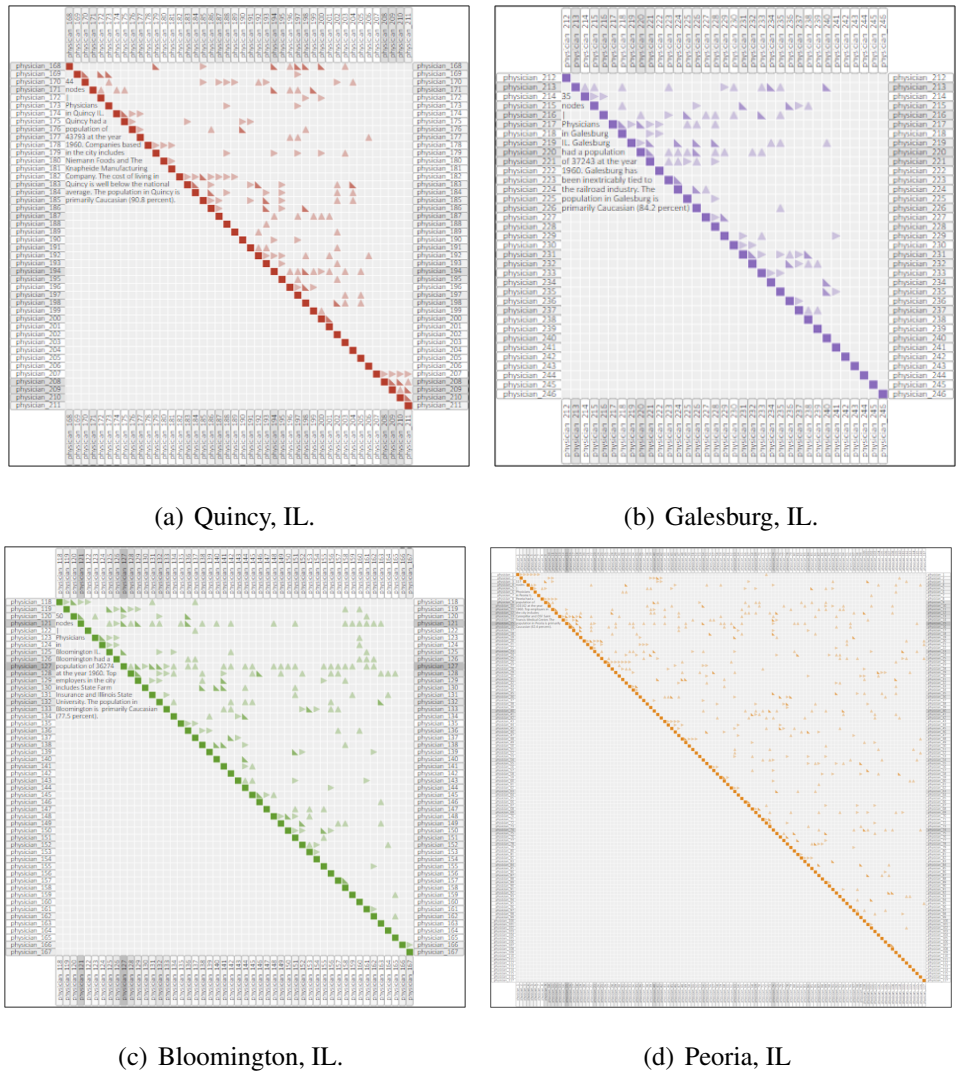
7.3.1 Data

This case study visualizes a non-weighted, directed, ground-truth community network for the diffusion of innovations between 246 physicians from Illinois in the year of 1966 [165]. Each node represents a physician. The dataset includes additional information on the physicians, such as the cities in which the physicians were practicing, their social circle formation, the length of time they have been practicing, the fields they specialize in, etc. This information can be used to group physicians into different communities. An edge connecting node E to node F indicates that Physician E considers reaching out to Physician F for the exchange of information.

This dataset contains 246 nodes, 1098 edges, and 13 possible community groupings. For this case study, I focus on two ground-truth communities grouped by the cities in which the physicians practiced, and the duration the physicians had been practicing in their corresponding cities.

7.3.2 Community Selection and External Context

Practicing City



(a) Quincy, IL. (b) Galesburg, IL.

(c) Bloomington, IL. (d) Peoria, IL

Fig. 7.16. Behavior comparison between physician networks in four cities.

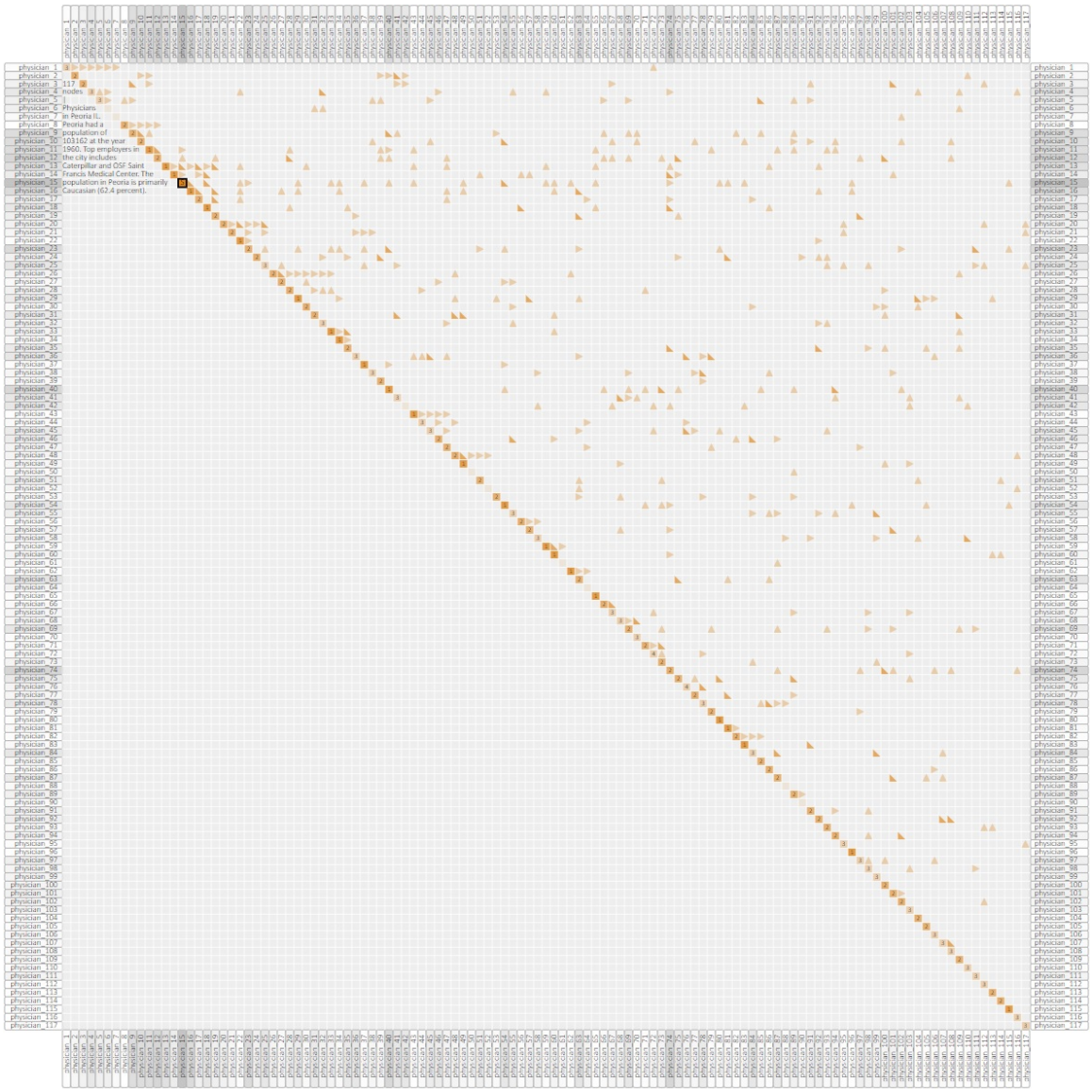


Fig. 7.17. Node 15 is highly influential in the Peoria community.

Fig. 7.15 visualizes the dataset with ground-truth communities based on the cities the physicians practiced. This dataset focuses on physicians that practiced at four Illinois cities: Peoria, Bloomington, Quincy, and Galesburg. As shown in the figure, the connections between the physicians are mostly restricted to their cities. While Peoria is the only independent community, the remaining communities only require minimal shifting of nodes to

create four independent communities. This indicates that geographical communities serve as a major factor in forming physician networks. However, as shown in Fig. 7.16, no obvious behavior differences between the four communities are found through comparing the density and edge types within each community.

With the additional context imported, as shown in Fig. 7.15, we can see the four cities have similar physician-to-population ratios, close to 1 to 1000. Based on the economy-related context, we can also hypothesize that the reason Peoria has a more independent physician network might be related to one of its top employers being a medical center, therefore allowing a more self-contained medical community. It is also interesting to observe that, while the populations in all four cities are primarily Caucasian, communities with a lower ratio of the Caucasian population have lower external edges.

Similarly to the previous case study, users can extract information such as influencers and neighboring distance to evaluate the diffusion of information or innovations for each city. For example, Fig. 7.17 shows physician 15 to be the influencer in the Peoria community. They can reach all but 9 out of the 117 physicians in the community. Out of all of the reachable nodes, only one node is at a degree-4 distance, while the remaining nodes are under degree-3 distances.

Practicing Duration

Fig. 7.18 visualizes the dataset with ground-truth communities based on how long the physicians practiced within their corresponding communities. This community selection separates the physicians into four groups: physicians that have been practicing for less than 10 years, physicians that have been practicing between 10 and 20 years, physicians that have been practicing for over 20 years, and physicians that did not report the duration of their practices. As shown in the figure, when switched to the duration-based grouping, the number of cross-community edges increases significantly, and the edges are evenly spread throughout the nodes in the communities. This suggests that practicing duration has little impact on the connections between physicians. While I originally predicted physicians

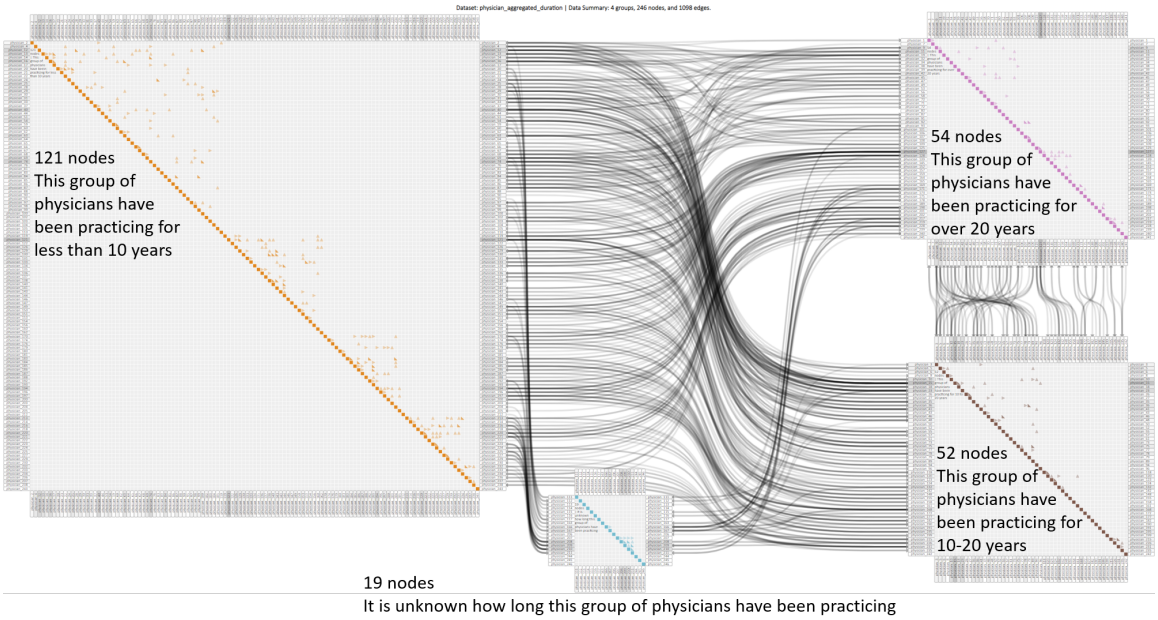
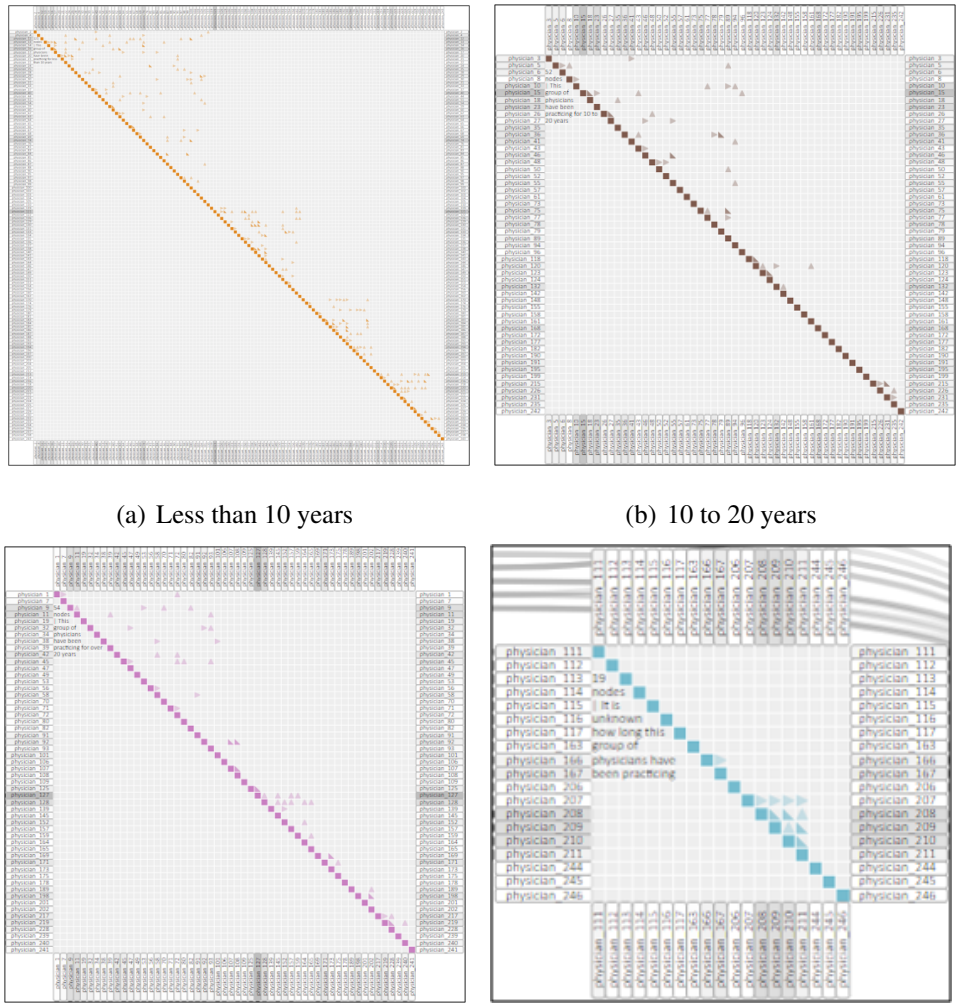


Fig. 7.18. A physician network, grouped by practicing duration, visualized using the Three-Component Visual Summary design.

would be more likely to reach out to physicians that are more senior in the field, the directional arrows connecting the different community matrices show no apparent difference between the two directions.

Fig. 7.19 shows the connections within each community to be noticeably more sparse compared to the city-based communities. Additionally, we can see each community contains multiple independent network subsets, which are likely based on the practicing cities. As with the city-based communities, there is no obvious difference between the network behavior within each community, which is observed through the edge patterns and the label shadings. Overall, grouping by practicing duration provides a less effective grouping for this Three-Component Visual Summary design, as it increases the number of crossing external edges significantly. However, each community selection provides unique insights into understanding the dataset.



(a) Less than 10 years

(b) 10 to 20 years

(c) Over 20 years

(d) Unknown duration

Fig. 7.19. Behavior comparison between physician networks based on practicing duration.

7.4 Feedback

This design was informally assessed in February 2020 by two University faculty members that are in decision-making positions and often interact with network data. Neither faculty were involved in the decision-maker survey described in Chapter 3.2. In this section, I summarize the feedback received from informal interviews with the decision-makers after demonstrating the system.

Faculty 1 first confirmed the challenge of not having the time and technical expertise to utilize the full potential of visual analytics tools. She described her processes of collecting the information needed for her decisions as relying primarily on her data analysts, who often present to her using static screenshots from visual analytics toolkits. Faculty 1 found this tool to be a great help to her work. She appreciated "the clarity and the preciseness of the display" and the way "the system reduces the workload required to trace the connections between nodes." She stated that the system allows her to gain a quick grasp of the dataset and is a tool she feels like she can use herself. She stated her "ideal scenario would be for the data analyst to present to [her] datasets using this system directly." In a similar amount of time as the usual presentation, she could use the tool to discuss with her analysts and update the selections in the system to examine the analysis results immediately, and store the screenshots for later use. However, she did express a tendency to identify additional interests when examining the visualization and wanted "an interactive way for the user to define new questions to ask or network parameters to subset on while looking at the data with the analyst."

Additionally, faculty 2 complimented "the approach of displaying patterns and data simultaneously." He also agreed with how "the visualization suggests an order in which to examine the three components" and "fades additional text to the background." He noted that it is "important to ensure continuity of visualization elements and clear hierarchy." While he liked the design idea of the Three-Component Visual Summary and recognized the value of the design in bridging the communication gap between data analysts and decision-makers, he urged the designer to "emphasize patterns (such as the flow of data) further", to "ensure clarity of each individual information element", and to "incorporate more automated features into the system" such as automatic sorting based on neighbor distance when a node is selected. Finally, he encouraged designers to "consider affordance theory [166,167] and the C.R.A.P Principles of Graphic Design [168] when selecting visual encoding for the three components." This aligns with the guideline to use conventional visualization techniques and signifiers to reduce the required technical expertise from the decision-makers.

Similarly to the feedback presented in the previous chapter, this summary is the result of informal discussions with domain experts on the usability and personal relevance of the system. Written notes were taken during the discussions, but no official reports were generated. The MILCs [158]-inspired questionnaire used to initiate the discussions can be found in Appendix B.

7.5 Discussion

Henry et al.'s NodeTrix design [14] established an effective display for the representative data and the data envelope. Interestingly, they also decided to improve network representation by providing both a readable global structure and the local communities in one visualization rather than the use of interactive exploration. The interaction functionalities provided in NodeTrix focus on building and rearranging the node-link and matrix combination to form a balanced display and create "meaningful summary visualizations of [analysts'] finding." This direction also aligns with the goal of this research work. However, the main challenge for my target audience, decision-makers who are casual experts, in adapting to NodeTrix's summary visualization is that using NodeTrix effectively still requires the skill and the time of a data analyst. While NodeTrix allows the analysts to arrange the dataset in a manner that is convenient to analyze, analysts still need to perform the network analysis the traditional way. Building on top of NodeTrix's design, the Three-Component Visual Summary design for network data inherits many of NodeTrix's strengths in effectively communicating the global structure and the community details of large scale network datasets, and adds to the system by inserting the analysis results into the display directly to reduce the time and skill required for decision-makers who are casual experts. With the network Three-Component Visual Summary, a decision-maker will not need to manually trace the connections within the matrices to identify paths connecting communities and nodes and so on, and can focus on identifying which relations to explore. This makes the best use of decision-makers' domain expertise and the computational power of machines. Since correlation does not always equal causation, the additional descriptions

added to each community can also help decision-makers tie the insights discovered to real-world context and better confirm or dismiss the findings.

While designed to handle large scale datasets, the network Three-Component Visual Summary still has a limit to its scalability. Unlike the designs of the three other applications, each data entry in the dataset is adequately visualized. This is necessary as many of the analyses focus on specific nodes, which means each node needs enough pixels in the display to allow proper communication of its identity and the related analyses and provide a recognizable and accessible area for click events. The design allows easier navigation of large datasets by reducing the number of crossing edges, but does not necessarily reduce the number of pixels needed to visualize each data entry. With how compact the adjacency matrix representation is, it is also difficult to overlay additional information on top of it, making the use of tooltip necessary. However, decision-makers may not find tooltips an appealing visual component in the final static presentation.

While this design does not consider weighted edges, as I am more interested in connections between players and would like to keep the initial design simple, it can easily adapt weight into the design. In the adjacency matrices, the weight of the edges can be encoded into the opacity of each cell. In the node-link diagram, the weight of the edges can be encoded into either the width or the opacity of the paths. Weight can also be incorporated into the calculation of neighbor distance and shortest path if applicable. Ultimately, the design should be customized to accommodate the needs of the user.

8. DISCUSSIONS AND RECOMMENDATIONS

This chapter discusses the collective insights gained through designing the Three-Component Visual Summary and the four applications. With these insights, I provide recommendations for applying the Three-Component Visual Summary design.

8.1 Strengths

Differing from summarization techniques focusing on the aggregation of data and equally weighted presentation of data entries, the Three-Component Visual Summary highlights components more relevant to the decisions to be made and dims the entries that contribute less to the decision. This approach prevents data entries of high importance from being accidentally dropped during the summarization process. With the different priorities reflected in their corresponding visual design, the visual design also provides audiences with a defined order to explore the visual summary in a manner similar to the Shneiderman mantra [2] without requiring interactive exploration, which many of the domain experts do not have the time or training for. The flexibility to choose the variables for the three components also allows the visual summary to be a more customized and focused experience, which works toward reducing the communication gap.

The Three-Component Visual Summary design also improves the storytelling of the data by balancing between author-driven and reader-driven stories and supporting the system with constrained interaction [7]. The balance between author-driven and reader-driven stories is achieved by pre-generating a selected list of relevant analyses and context and allowing users to select the desired combination. Constrained interaction is used to support the selection of components and highlighting attributes of the data entries. With the goal of communicating more comprehensive storytelling of data and analysis results while allowing simple backtracking and reasoning, this design falls between the exploratory and

confirmatory spectra of visualization, leaning slightly toward confirmatory as specific results are highlighted. However, designers can adjust the number of options allowed to alter the display in order to shift a design toward one spectrum. The more options allowed, the more exploratory a design will be, and vice versa. The improvement in storytelling is reflected by the preference of the domain experts and prospective users for using designs presented in this dissertation over their traditional practices.

The strengths of this work are reflected in the evaluation results of the four designs. From the numerical design and its study, we can see the Three-Component Visual Summary design increases the accuracy in performing decision-making inspired tasks when tested against common visualization techniques (Chapter 4.2). While it may seem obvious, having a design that focuses on the needs of the audience allows the user to perform better than using a generalized tool. From the contextual design, we see the Three-Component Visual Summary design creates a coherent display of the dataset and generated a more preferred way of storytelling (Chapter 5.4.4). From the geospatial design, we see the Three-Component Visual Summary design is considered to be accessible to most decision-makers, and the inclusion of external context is considered a welcoming addition that is not utilized enough in the current practices (Chapter 6.3). From the network design, we see the Three-Component Visual Summary design increases the precision of the presentation (Chapter 7.4). It also allows decision-makers with limited time and technical expertise to properly utilize the power of visual analytics. With the Three-Component Visual Summary design proved effective across the four major data types utilized in commercial visual analytics tools, I believe this design can benefit decision-makers in most domains.

8.2 Limitations

While the customizable aspect of the Three-Component Visual Summary design reduces the difficulty for casual experts to make use of visualizations, the outcome still depends on the designers' ability to select the appropriate variables and the corresponding visual representations to construct effective three-component visual summaries. Even though

there exist guidelines and tools for identifying visualization techniques suitable for different data types (e.g., The Grammar of Graphics [169], The TimeVis Browser [1], and Text Visualization Browser [2]), in-depth studies on combining and connecting different visualization techniques into one static display are lacking. Additionally, using the customized visual summary instead of a generalized tool for presentation means that spontaneous analysis that is not incorporated into the design and more sophisticated analyses may not be achievable through the Three-Component Visual Summary design. From the feedback received, we know the design is effective in identifying "what you know that you do not know", but it is often valuable to be able to identify "what you do not know that you do not know" as well. The discovery of "unknown unknowns" is not well incorporated into the Three-Component Visual Summary, fundamentally because of the design direction in automatically calculating and highlighting the "known unknowns". However, this tool is not meant to replace traditional visual analytics, but rather serve as an additional tool to reduce the communication gap between the data analysts and the decision-makers. This highlights the need for designers to be able to quickly update the Three-Component Visual Summary to incorporate a new analysis of interest.

Even though the goal of the Three-Component Visual Summary design is to allow domain experts with little to no training in visual analytics to be able to make use of data visualization, it remains difficult to create a design that can guarantee casual experts will be able to use it without any previous skill in visualization. The design tries to minimize the training necessary by utilizing simple visual designs, familiar signifiers, and provide instructions throughout the system. It is recommended to consider both affordance theory [166, 167] and the C.R.A.P Principles of Graphic Design [168] in the design process to achieve the same goal. The design also focuses on reducing the need for interactive exploration to reduce the technical expertise necessary to use the system and allows users to focus on learning the visualization. While it is difficult to eliminate training completely, the result from the quantitative study in Chapter 4 shows that minimal training is sufficient.

[1] http://browser.timeviz.net/

[2] https://textvis.lnu.se/

8.3 Design Comparisons and Discussions

When constructing the Three-Component Visual Summary designs for the four data types, different approaches were taken because of the differences in data characteristics and visualization techniques available. This section discusses the different approaches and their impact.

Visualizing the analysis results directly versus indirectly: In Chapter 4, the summarized line graph utilizes the highlighted and aligned extrema to support the estimation of multiple analyses relevant to the decisions. Extracting analysis results from the supporting tool that is visualized on top of the data envelope component allows users to gain an understanding of "why" without the need for an interactive system. Compared to the numerical design, the geospatial design and the network designs allow users to select the analyses of interest to be displayed directly in the visual summary generated, rather than utilizing a supporting tool to aid the extraction of analysis results. In these two cases, the "why" is provided through the layering of other visual components in the geospatial design and through the additional context and the ability to trace the path in the network design. While both approaches help users make sense of the analysis results, the first approach can embed more analyses in a static visual summary, and the second approach allows more direct and efficient extraction of the analysis results. This means that while the improvement in completion time from the summarized line graph user study (Chapter 4.2) was not statistically significant, the geospatial design and the network design should allow a quicker generation of relevant insights that is more significant without losing the context. Similarly, the alternate numerical design described in Chapter 4.1.5 should allow greater improvements in completion time, although the design may require adjustments in scaling to incorporate the context clearer.

Connecting the three components through an direct overlay approach versus an interconnected multi-view approach: Of the four applications, we can see that the numerical and the geospatial designs are the most suitable for a direct overlay approach when combining the three components as the three components are able to find a shared ba-

sis to be drawn upon. For the numerical application, all three components can share the same x-axis and y-axis. For the geospatial application, all three components can be drawn over the same geospatial coordinate system. The network application technically uses a direct overlay approach as well, with the three components sharing the same canvas. The main difference is that while the analytical highlights are layered on top of the NodeTrix-inspired diagram, the representative data and the data envelope are separated by a hierarchical structure instead of two structurally separated visual components. For the contextual application, the analytical highlights are gained through the visualization component that quantified the important metrics in the dataset, while the representative data utilize a text-based summary and the data envelope provides context through the original multimedia files. This makes connecting the three components through overlaying them directly while retaining the property to be able to generate effective static summary reports difficult. Instead of the direct overlay approach, the contextual application goes back to a multi-view approach. With the interconnected setup, the three views for the three components should be telling a connected story at all times. However, this approach still increases the effort required to connect the knowledge gained from the three different components.

Visualizing the individual data entries versus the aggregated data: Another separation between the four applications is whether the data entries are individually displayed or aggregated. In the contextual and geospatial applications, the data entries are being visualized in an aggregated manner. The geospatial application focuses more on collective boundaries and hotspots of data entries that belong to a selected category, and highlights specific locations rather than specific data entries. The contextual application quantifies the information extracted from the individual data entries to allow more sophisticated analyses and focuses on the dynamics and comparisons between the scale of the quantified metrics rather than a specific file or text. As a result, the two applications scale well with larger datasets, with the main restriction being the processing power given. The limitation in scalability, instead, is reflected in the number of spatial highlights displayed simultaneously for the geospatial design, and the number of metrics being compared simultaneously for the contextual application. On the other hand, both the numerical and the network application

visualize every data entry. The network design visualizes every node and edge in the dataset but minimizes the clutter of the edges within the communities by replacing the node-link diagram with an adjacency matrix representation. The numerical design dims out the original time-series and moves them to the background while highlighting two specific points from each time-series. While the numerical and network applications improve the scalability from the common line chart and node-link diagram, the scalability remains limited by the size of the dataset. However, there are other factors that can affect the scalability of the visual summaries. For example, the numerical design scales better with a larger dataset when the time-series have stronger correlations, as the contrast between the time steps of the maxima and the minima becomes more apparent. The network design also scales better with a larger dataset when the majority of edges are restricted within the communities as it would reduce the number of crossing edges.

Considering the designs generate visual summaries, it can be interesting to analyze the data loss of the Three-Component Visual Summary designs and compare the results between the four applications. The resolution loss and the variable loss of a visual summary can potentially be evaluated using information theory [170]. However, the research work for estimating information communicated over visualizations using information theory is still in an early stage. There is currently no well-established calculation method for precise measurement of data-loss. While high-level estimations can be made by comparing the differences in alphabet compression, potential distortion, and cost throughout statistics, algorithm, visualization, and interaction, the estimations vary greatly based on the specific visualization designs, the datasets, and the display spaces [171] and may not properly reflect the general effect of the three-component design. The designers can, however, consider high-level principles (e.g., overlapping causes uncertainty, binning reduces precision, uncertainty leads to higher privacy but lowers utility [172]) when constructing a three-component design. It is also worth noting that information communicated over the visualization differs from the knowledge gained from the visualization. Higher display space utilization or visual mapping ratio does not necessarily result in higher utility or quality. Rather, the goal of a visual summary design should be to ensure the design loses infor-

mation optimally such that users can easily locate the information of interest. Given the research problem the Three-Component Visual Summary is designed for, data loss analysis may not be a critical metric. With the goal of bringing forth the data entries relevant to the decision to be made, and dimming out the remaining data entries to the background, it is more important for the Three-Component Visual Summary designs to retain and highlight the appropriate data entries efficiently, rather than simply minimizing the data loss in the graphic. On the other hand, knowledge can compensate for information loss [173], and the targeted audience of Three-Component Visual Summary designs are decision-makers who are also domain experts. Additionally, in the target scenario of the Three-Component Visual Summary design, the data analysts will still have access to the raw data, and the decision-makers will likely have limited time to focus on the highlighted items only. As a result, the impact of data loss analysis should be minimal.

8.4 Recommendations

The Three-Component Visual Summary can be applied to more datasets as long as the designer can identify components that satisfy the characteristics described in Chapter 3. In this section, I provide a list of recommendations based on what I learned from designing the visual summaries for the four applications.

- **Identify the characteristics to be compared**: It is important first to understand what characteristics in the dataset are to be compared to support the decision making. Knowing the characteristics to be compared can help the designer separate data entries that can be aggregated and data entries that cannot. This then allows the designer to identify the appropriate visual encoding. For example, the goal in the geospatial design is to compare the locations of data under different categories. As a result, the selected visual elements display aggregated attributes of the selected data subset on a fixed coordinate system and do not display the locations and content of the individual data entries.

- **Identify visualization techniques that share the same visualization basis for the three components**: While users can make the mental connection between the three components through the direct feedback of the multi-view approach, connecting the three components through a direct overlay approach reduces the dependency on an interactive system and supports more effective static visual summaries. For example, the data size, time, and location for the temporal analysis in the geospatial application can be encoded with an annotated line graph where the x-axis encodes the time, the y-axis encodes the volume of data within that temporal bin, and the annotated text describes the location by township or nearby landmarks. This visual encoding can allow more accurate visual comparison on data volume over time between multiple categories but will require a separate view from the data distribution visualization, which makes connecting and comparing findings from the two views more difficult. Therefore, when exploring different visual encodings fitting for the three components, it can be useful to identify visualization techniques sharing the same visualization basis first and prioritize experimenting with corresponding combinations.

- **Keep the visual representations of the three components in different styles**: It is best to use distinct visual styles to allow for a more distinct separation between the three components in a direct overlay design. For example, in the numerical application, the representative data uses a line, the analytical highlights are supported by small icons, and the data envelope utilizes semi-transparent areas. The three components can also be separated by different opacity levels or locations to suggest to users the examination order.

- **Provide context outside of data**: While many insights can be supported by the data itself, designers can link to external resources to provide additional context. In the geospatial application, landscape and census data are imported to provide additional context that may not be included in the raw data and are linked to the raw data through a shared coordinate system. A numerical data use case with stock market prices

can also import finance news on the companies included in the dataset to explain anomalous behaviors.

- **Understand the scalability limitations**: The scalability of a design may vary based on the data type and the visual encoding. Understanding the scale of each variable in the dataset, the possible visual encodings for those variables, and how those visual encodings can create visual clutter allows the designers to create more usable visual summaries. For example, knowing that visual clutter in a node-link diagram can be reduced when there are fewer crossing edges, a designer may be able to increase the scalability of the network visual summary by arranging the communities in a layout that has fewer crossing edges or selecting communities that have less external edges.

- **Utilize common signifiers and combine conventional visualization techniques**: To address the skill-level of casual experts, sometimes it is more effective to use standard visualization techniques that are simple than powerful but new visualization techniques. Supporting the constrained interactions with familiar signifiers and incorporating affordance theory [166, 167] and the C.R.A.P. Principles [168] into the designs also reduces the learning curve.

- **Use additional text or figures to tell the story**: It is okay to overlay additional text or figures to the visual design to annotate the graphics and tell a stronger data story as long as they are not interfering with other crucial visual components. This practice is commonly seen in early data graphics such as those by William Playfair [20] before the support of interactive functionalities.

8.5 Summary Reference

Table 8.1 summarizes the visualizations explored by this thesis work and their compatibility with the Three-Component Visual Summary design. Specifically, this table presents visualization techniques or visual components that are appropriate for the four data types. For each visualization technique, the table highlights which of the three components it is

appropriate for, the information that can be encoded, the visualization basis, and the visualization style. The table separates the visualization techniques into the different components based on their abilities to convey a simple overview (RD: Representative Data), perform specific analyses (AH: Analytical Highlights), and present an aggregated summary of the raw data (DE: Data Envelope). The visualization style includes how the data is presented (individual, aggregated) and in what style it is visualized (point, glyph, line, area, view), which ties back to the recommendations. Based on the data types and the attributes a designer wants to present in a Three-Component Visual summary, the designer can identify first from the table which visualization techniques are capable of encoding the information desired. With those techniques, the designer can then explore the different combinations and attempt to construct a design that covers all three components, visualizes the components over the same basis, and utilizes different styles for each component. This table serves as a quick reference that complements the recommendations. Note that the analytical highlights component for the contextual application depends greatly on the important metrics of the quantified data. Also, most network visualization techniques can serve as the data envelope component by visualizing every node and edge. The separation between being categorized as appropriate representative data or analytical highlights lies in whether it is more effective at communicating the overall network structure or specific attributes.

Table 8.1.
A summary of the visualizations explored and a reference for the recommendations.

	Component	Encoding	Basis	Style	Visualization
Numerical	RD, AH	Mean, Median, Best Fit, Statistical Analyses	Specified 2D Plane	Individual, Line	Line Graphs
	RD	Proportion Overview	Abstract 2D Plane	Aggregated, Area	Pie Charts
	AH	Individual and Overall Trends	Abstract 2D Plane	Individual, View	Horizon Graphs
	AH	Thresholds, Trends, Alignments	Specified 2D Plane	Individual, Line	(Straight) Lines
	AH, DE	Statistical Analyses, Progress, Count, Value	Specified 2D Plane (Along One Axis)	Either, Area	Bar Charts
	AH, DE	Overall Trend, Raw Data	Specified 2D Plane	Aggregated, Area	Stream Graphs
	DE	Raw Data, Distributions	Specified 2D Plane	Aggregated, Area	Density Bands
	DE	Overall Range	Specified 2D Plane	Aggregated, Area	Band Graphs
	DE	Raw Data	Specified 2D Plane	Individual, Point	Scatter Plots
	DE	Hierarchical Sizes and Trends	Abstract 2D Plane	Aggregated, Area	Tree Maps
Contextual	RD	Text Summary, Takeaway Values	Abstract 2D Plane	Aggregated, Area	Short Text
	RD	Overall Progress Percentage	Abstract 2D Plane	Aggregated, Area	Liquid Fill Gauge/Radio Progress Bar
	AH	Quantified Comparisions	Depends	Depends	Quantified Visual Representations
	DE	Events	Abstract 2D Plane	Aggregated, Area	Storyboard
	DE	Tags, Topics, Metadata	Abstract 2D Plane	Aggregated, Area	Word Cloud
	DE	Original File(s)	Existing Elements	Depends	Hyperlink
	DE	Detailed descriptions	Abstract 2D Plan	Aggregiated, Area	Description Text
Geospatial	RD, AH	Range/Boundary, Clusters	Geospatial 2D Plane	Individual, Line	Contour
	AH	Regional Statistical Analyses	Geospatial 2D Plane	Aggregated, Area	Choropleth Map
	AH	Regional Statistical Analyses	Geospatial 2D Plane	Individual, Area	Charts
	AH	Temporal Movements	Geospatial 2D Plane	Individual, Line	Trajectories
	AH, DE	Distributions	Geospatial 2D Plane	Aggregated, Area	Heatmaps
	DE	External Geographical Context	Geospatial 2D Plane	External, Background	Map Tiles
	DE	Raw Data	Geospatial 2D Plane	Individual, Point	Points
Network	RD, DE	Network Connections, Network Structure	Abstract 2D Plane	Individual, Area	Node-Link Diagram
	RD, DE	Network Connections, Network Structure	Abstract 2D Plane (Radial Layout)	Individual, Area	Hive Plots
	AH	Communities	Abstract 2D Plane	Aggregated, Lines	Contour Boundaries
	AH, DE	Network Connections	Abstract 2D Plane	Individual, Area	Adjacency Matrices
	AH, DE	Network Connections, Flow Proportion	Abstract 2D Plane	Individual, Area	Sankey Diagrams
	AH, DE	Network Connections, Flow Proportion	Abstract 2D Plane (Radio Layout)	Individual, Area	Chord Diagrams
	DE	Network Connections	Abstract 2D Plane (Along One Axis)	Individual, Area	ARC Diagrams
Universal	AH	Correlation, Group, Scale, Outliers	Existing Elements	Depends	Color/Texture Variations
	AH	Outliers, Extrema, Events	Specified 2D Plane	Individual, Glyph	Icons
	AH, DE	Analyses, Context	Abstract 2D Plane	Individual, Area	Annotations

9. CONCLUSIONS

The main goal of this dissertation is to provide a solution to the challenges that keeps casual experts from fully utilizing the power of visual analytics to support data-driven decisions. Specifically, the solution needs to address the communication gap between data analysts and decision-makers. In this chapter, I summarize my thesis work and explain the next steps to move this research forward.

The first part of this research focused on gaining an in-depth understanding of the challenges and designing a solution accordingly. I surveyed six decision-makers employed by safety agencies to gather an understanding of the interactions between decision-makers and data analysts (Chapter 3.2). A list of design requirements for the solution was then derived from the survey findings (Chapter 3.1). Following the design requirements, I proposed a Three-Component Visual Summary design that combines a high-level overview, analytical highlights and comparisons, and context relevant to the analyses into a visual summary that allows efficient and accurate knowledge extractions and validations (Chapter 3.3). This design aims to reduce the time and technical expertise required for visual analytics tools by highlighting the known decision-relevant analysis results and incorporate the knowledge traditionally gained from different levels of exploration into the visual design. This design also aims to bridge the communication gap connecting the highlighted analyses to the corresponding context to reduce the impact of the potential information bias.

The second part of this research focused on evaluating and understanding how applicable and effective the solution is and answering questions such as: “Does the solution provide significant enough improvements?” “Can the solution be applied to different datasets?” I evaluated the design by applying it to four datasets, from which four new visualizations were created, each using one of the four major data types used in commercial visual analytics tools. The numerical application visualized stock market data and improved the accuracy in decision-making inspired tasks (Chapter 4.2). The contextual application

visualized the impact of an academic research group and was the preferred way of telling the story behind the data (Chapter 5.4.4). The geospatial application visualized crime reports and provided a new perspective into exploring the crime report data for data-driven decision making (Chapter 6.3). The network application visualized data flow networks and allowed the decision-maker to have a quicker and more precise grasp of the dataset (Chapter 7.4). The evaluation results suggested that the design is applicable to a wide range of datasets and produces consistent improvements to the understanding and the communication of the data. Based on the insights gained through extracting the design requirements and applying the design to the different datasets, I compiled a list of recommendations on implementing the Three-Component Visual Summary design.

This dissertation marks a first step in designing decision-driven visualizations for casual experts. With the solution proposed in this thesis validated in experimental and practical settings, future work should, naturally, focus on making the solution accessible and easy to implement to generate more practical (in contrast to theoretical or empirical) contributions. Below is a list of future directions identified from the feedback and the discussions to continue moving this research work forward:

- **Survey common data types, critical decisions, and relevant analyses for decision-makers in different fields**: Having a reference on the visualization techniques suitable for different decision-makers can reduce the effort required to design a Three-Component Visual Summary. Chapter 8 provides a simple reference for selecting the appropriate visualization techniques to combine and effectively communicate the analysis results based on the different data types. This reference basis is appropriate for the scope of this thesis as one of the goals is to understand the applicability of the solution that is the Three-Component Visual Summary across different data types. However, when designing a visual summary to communicate to a specific decision-maker, the reference and the recommendation should be based on the domain of the decision-maker. A more in-depth understanding of the standard data collected, critical decisions that are often made, and analyses that can generate insights relevant to those decisions for the different fields will greatly expand the reference and support a

more fluid and customized design process focusing on the decision-maker. Building such a reference will require extensive surveys of decision-makers at different levels and in various fields of work.

- **Explore and define the appropriate evaluation metrics for combined visualization techniques**: While guidance, surveys, and research work on the strengths and trade-offs of different visualization techniques exist, the overall effectiveness of a visualization that consists of multiple connected techniques is not a simple union or intersection of the strengths and weaknesses from the individual techniques. Metrics used to evaluate the usability of the design, such as scalability, will have to be measured differently. For example, the scalability of my proposed geospatial design in Chapter 6 is more dependent on the number of visual components displayed in the same visual style than it is the least scalable visual component incorporated or the number of data entries. With the number of possible visualization techniques to incorporate into the design increasing immensely because of the combination aspect, a set of evaluation metrics on the overall effectiveness of the Three-Component Visual Summary designs can help designers rank the different combinations of visualization techniques based on the scenario and identify the most fitting visual summary design. Having such impact metrics defined can also help to establish a more structured design process for the Three-Component Visual Summary design.

- **Create a tool to support creating and/or updating Three-Component Visual Summary designs**: Based on the feedback from prospective users (described in Chapter 6.3 and Chapter 7.4), it is not uncommon for a decision-maker to discover a new analysis of interest when examining the visual summaries. If the analysis of interest is unexpected and not previously incorporated into the design, the designer needs to be able to update the design quickly to provide the needed information. However, incorporating a new visual component for a new analysis into an established design can damage the visual balance between the existing visual components. It is important for designers to be able to quickly construct or update an effective

Three-Component Visual Summary design for this solution to be practical. With a domain-focused reference built and a set of evaluation metrics defined, the final step to make designing a Three-Component Visual Summary more straightforward and efficient will be creating an authoring system for the Three-Component Visual Summary design. This tool should provide a list of recommended visualization techniques based on the decision and the domain of the target audience, evaluate the combinations of the selected techniques, and construct the Three-Component Visual Summary with the data provided.

REFERENCES

REFERENCES

[1] W. S. Cleveland, *Visualizing Data*. Hobart Press, 1993.

[2] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," in *Proceedings of the IEEE Symposium on Visual Languages*. IEEE, 1996, pp. 336–343. doi: 10.1109/VL.1996.545307

[3] K. Madhavan, N. Elmqvist, M. Vorvoreanu, X. Chen, Y. Wong, H. Xian, Z. Dong, and A. Johri, "DIA2: Web-based cyberinfrastructure for visual analysis of funding portfolios," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1823–1832, Dec 2014. doi: 10.1109/TVCG.2014.2346747

[4] A. Sarikaya, M. Gleicher, and D. Szafir, "Design factors for summary visualization in visual analytics," *Computer Graphics Forum*, vol. 37, pp. 145–156, 2018. doi: 10.1111/cgf.13408

[5] F. B. Viegas and M. Wattenberg, "Communication-minded visualization: A call to action," *IBM Systems Journal*, vol. 45, no. 4, pp. 801–812, 2006. doi: 10.1147/sj.454.0801

[6] Z. Pousman, J. Stasko, and M. Mateas, "Casual information visualization: Depictions of data in everyday life," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1145–1152, Nov 2007. doi: 10.1109/TVCG.2007.70541

[7] E. Segel and J. Heer, "Narrative visualization: Telling stories with data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1139–1148, Nov 2010. doi: 10.1109/TVCG.2010.179

[8] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context preserving, dynamic word cloud visualization," in *Proceedings of the IEEE Pacific Visualization Symposium,*. IEEE, 2010, pp. 121–128. doi: 10.1109/MCG.2010.102

[9] J. Kruger, J. Schneider, and R. Westermann, "Clearview: An interactive context preserving hotspot visualization technique," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 941–948, 2006. doi: 10.1109/TVCG.2006.124

[10] W. Javed, S. Ghani, and N. Elmqvist, "Polyzoom: Multiscale and multifocus exploration in 2d visual spaces," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 287–296. doi: 10.1145/2207676.2207716

[11] W. Javed and N. Elmqvist, "Stack zooming for multi-focus interaction in time-series data visualization," in *Proceedings of the IEEE Pacific Visualization Symposium*. IEEE, 2010, pp. 33–40. doi: 10.1109/PACIFICVIS.2010.5429613

[12] W. Javed and N. Elmqvist, "Stack zooming for multifocus interaction in skewed-aspect visual spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 8, pp. 1362–1374, 2013. doi: 10.1109/TVCG.2012.323

[13] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstadt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. Keim, "Visual analytics for the big data era—a comparative review of state-of-the-art commercial systems," in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2012, pp. 173–182. doi: 10.1109/VAST.2012.6400554

[14] N. Henry, J.-D. Fekete, and M. J. McGuffin, "NodeTrix: a hybrid visualization of social networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1302–1309, 2007. doi: 10.1109/TVCG.2007.70582

[15] P. Mann, *Introductory Statistics*. John Wiley & Sons, 2010.

[16] A. Bryman and D. Cramer, *Quantitative Data Analysis for Social Scientists*. New York, NY, 10001: Routledge, 1994.

[17] M. F. De Oliveira and H. Levkowitz, "From visual data exploration to visual data mining: a survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 3, pp. 378–394, July 2003. doi: 10.1109/TVCG.2003.1207445

[18] P. C. Wong and J. Thomas, "Visual analytics," *IEEE Computer Graphics and Applications*, vol. 24, no. 5, pp. 20–21, Sept 2004. doi: 10.1109/MCG.2004.39

[19] M. Correll, D. Albers, S. Franconeri, and M. Gleicher, "Comparing averages in time series data," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1095–1104. doi: 10.1145/2207676.2208556

[20] W. Playfair, *The Commercial and Political Atlas: Representing, by Means of Stained Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure and Debts of England During the Whole of the Eighteenth Century*. T. Burton, 1801.

[21] W. Playfair, *The Statistical Breviary: Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe*. T. Bensley, Bolt Court, Fleet Street, 1801.

[22] B. Shneiderman, "Tree visualization with tree-maps: 2-D space-filling approach," *ACM Transactions on Graphics*, vol. 11, no. 1, pp. 92–99, Jan. 1992. doi: 10.1145/102377.115768

[23] N. Elmqvist and J.-D. Fekete, "Hierarchical aggregation for information visualization: Overview, techniques and design guidelines," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 439–454, May 2010. doi: 10.1109/TVCG.2009.84

[24] J. Heer and M. Agrawala, "Multi-scale banking to 45 degrees," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 701–708, 2006. doi: 10.1109/TVCG.2006.163

[25] J. Heer and B. Shneiderman, "Interactive dynamics for visual analysis," *Queue*, vol. 10, no. 2, p. 30, 2012. doi: 10.1145/2133806.2133821

[26] S. M. Kocherlakota and C. G. Healey, "Interactive visual summarization of multidimensional data," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2009, pp. 362–369. doi: 10.1109/ICSMC.2009.5346206

[27] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context preserving dynamic word cloud visualization," in *Proceedings of the IEEE Pacific Visualization Symposium*. IEEE, 2010, pp. 121–128. doi: 10.1109/MCG.2010.102

[28] A. Perrot, R. Bourqui, N. Hanusse, F. Lalanne, and D. Auber, "Large interactive visualization of density functions on big data infrastructure," in *Proceedings of the IEEE Symposium on Large Data Analysis and Visualization*. IEEE, Oct 2015, pp. 99–106. doi: 10.1109/LDAV.2015.7348077

[29] N. Elmqvist, N. Henry, Y. Riche, and J.-D. Fekete, "Mélange: Space folding for visual exploration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 468–483, 2010. doi: 10.1109/TVCG.2009.86

[30] Q. W. Bouts, T. Dwyer, J. Dykes, B. Speckmann, S. Goodwin, N. H. Riche, S. Carpendale, and A. Liebman, "Visual encoding of dissimilarity data via topology-preserving map deformation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 9, pp. 2200–2213, 2016. doi: 10.1109/TVCG.2015.2500225

[31] S. M. Kocherlakota and C. G. Healey, "Interactive visual summarization of multidimensional data," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. IEEE, Oct 2009, pp. 362–369. doi: 10.1109/ICSMC.2009.5346206

[32] J. Hullman and N. Diakopoulos, "Visualization rhetoric: Framing effects in narrative visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2231–2240, 2011. doi: 10.1109/TVCG.2011.255

[33] "Many eyes," http://www.boostlabs.com/ibms-many-eyes-online-data-visualization-tool/, accessed: 2017-06-14.

[34] C. Bryan, K.-L. Ma, and J. Woodring, "Temporal summary images: An approach to narrative visualization via interactive annotation generation and placement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 511–520, 2017. doi: 10.1109/TVCG.2016.2598876

[35] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: automatic generation of annotated stock visualizations," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2707–2716. doi: 10.1145/2470654.2481374

[36] M. Friendly and D. Denis, "The early origins and development of the scatterplot," *Journal of the History of the Behavioral Sciences*, vol. 41, no. 2, pp. 103–130, 2005. doi: 10.1002/jhbs.20078

[37] L. Byron and M. Wattenberg, "Stacked graphs—geometry & aesthetics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1245–1252, Nov 2008. doi: 10.1109/TVCG.2008.166

[38] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, no. 2, pp. 69–91, 1985. doi: 10.1007/BF01898350

[39] D. F. Williamson, R. A. Parker, and J. S. Kendrick, "The box plot: a simple visual method to interpret data," *Annals of Internal Medicine*, vol. 110, no. 11, pp. 916–921, 1989. doi: 10.7326/0003-4819-110-11-916

[40] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploratory spatio-temporal visualization: an analytical review," *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 503–541, 2003. doi: 10.1016/S1045-926X(03)00046-6

[41] E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose, "Toolglass and magic lenses: the see-through interface," in *Proceedings of the Conference on Computer graphics and interactive techniques*. ACM, 1993, pp. 73–80. doi: 10.1145/259963.260447

[42] A. Slingsby, J. Dykes, J. Wood, and K. Clarke, "Interactive tag maps and tag clouds for the multiscale exploration of large spatio-temporal datasets," in *Proceedings of the International Conference Information Visualization*. IEEE, 2007, pp. 497–504. doi: 10.1109/IV.2007.71

[43] S. Afzal, R. Maciejewski, Y. Jang, N. Elmqvist, and D. S. Ebert, "Spatial text visualization using automatic typographic maps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2556–2564, 2012. doi: 10.1109/TVCG.2012.264

[44] A. Malik, R. Maciejewski, T. F. Collins, and D. S. Ebert, "Visual analytics law enforcement toolkit," in *Proceedings of the IEEE International Conference on Technologies for Homeland Security*. IEEE, 2010, pp. 222–228. doi: 10.1109/THS.2010.5655057

[45] A. Malik, R. Maciejewski, B. Maule, and D. S. Ebert, "A visual analytics process for maritime resource allocation and risk assessment," in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2011, pp. 221–230. doi: 10.1109/VAST.2011.6102460

[46] C. Gorg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko, "Visual analytics with Jigsaw," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2007, pp. 201–202. doi: 10.1109/VAST.2007.4389017

[47] S. Havre, B. Hetzler, and L. Nowell, "Themeriver: Visualizing theme changes over time," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2000, pp. 115–123. doi: 10.1109/INFVIS.2000.885098

[48] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "Textflow: Towards better understanding of evolving topics in text," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412–2421, 2011. doi: 10.1109/TVCG.2011.239

[49] C. Collins, F. B. Viegas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009, pp. 91–98. doi: 10.1109/VAST.2009.5333443

[50] J. Zhang, J. Chae, S. Afzal, A. Malik, D. Thom, Y. Jang, T. Ertl, S. A. Matei, and D. S. Ebert, "Visual analytics of user influence and location-based social networks," in *Transparency in Social Media*. Springer, 2015, pp. 223–237. doi: 10.1007/978-3-319-18552-1_12

[51] K. Shibata, A. Takahashi, and T. Shirai, "Fault diagnosis of rotating machinery through visualisation of sound signals," *Mechanical Systems and Signal Processing*, vol. 14, no. 2, pp. 229–241, 2000. doi: 10.1006/mssp.1999.1255

[52] T. Gorko, C. Yau, A. Malik, M. Harris, J. X. Tee, R. Maciejewski, C. Qian, S. Afzal, B. Pijanowski, and D. Ebert, "A multi-scale correlative approach for crowd-sourced multi-variate spatiotemporal data," in *Proceedings of the Hawaii International Conference on System Sciences*. University of Hawaii at Manoa, 2018. doi: 10.24251/HICSS.2018.213

[53] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 296–305, Feb 2005. doi: 10.1109/TCSVT.2004.841694

[54] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proceedings of the ACM Conference on Multimedia*. New York, NY, USA: ACM, 2002, pp. 533–542. doi: 10.1145/641007.641116

[55] J. Zhang, J. Chae, C. Surakitbanharn, and D. S. Ebert, "SMART: Social media analytics and reporting toolkit," in *Proceedings of the IEEE Workshop on Visualization in Practice*. IEEE, 2017, pp. 1–5.

[56] E. Tufte, *The Visual Display of Quantitative Information*. Graphics Press, 1983. doi: 10.1097/01445442-198507000-00012

[57] E. W. Anderson, K. C. Potter, L. E. Matzen, J. F. Shepherd, G. A. Preston, and C. T. Silva, "A user study of visualization effectiveness using EEG and cognitive load," *Computer Graphics Forum*, vol. 30, no. 3, pp. 791–800, 2011. doi: 10.1111/j.1467-8659.2011.01928.x

[58] H. Hochheiser and B. Shneiderman, "Interactive exploration of time series data," in *The Craft of Information Visualization*. Elsevier, 2003, pp. 313–315. doi: 10.1016/B978-155860915-0/50039-1

[59] R. Kincaid and H. Lam, "Line graph explorer: scalable display of line graphs using focus+ context," in *Proceedings of the ACM Conference on Advanced Visual Interfaces*. ACM, 2006, pp. 404–411. doi: 10.1145/1133265.1133348

[60] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom, "Visually mining and monitoring massive time series," in *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*. ACM, 2004, pp. 460–469. doi: 10.1145/1014052.1014104

[61] P. McLachlan, T. Munzner, E. Koutsofios, and S. North, "LiveRAC: interactive visual exploration of system management time-series data," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 1483–1492. doi: 10.1145/1357054.1357286

[62] J. J. Van Wijk and E. R. Van Selow, "Cluster and calendar based visualization of time series data," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 1999, pp. 4–9. doi: 10.1109/INFVIS.1999.801851

[63] M. Wattenberg, "Visualizing the stock market," in *Proceedings of the Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1999, pp. 188–189. doi: 10.1145/632716.632834

[64] R. Mundigl, "An underrated chart type: The band chart," http://www.clearlyandsimply.com/clearly_and_simply/2011/04/an-underrated-chart-type-the-band-chart.html, 2011, (Accessed 2018-02-28).

[65] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner, "Hierarchical parallel coordinates for exploration of large datasets," in *Proceedings of the IEEE Conference on Visualization*, IEEE Computer Society Press. IEEE, Oct 1999, pp. 43–50. doi: 10.1109/VISUAL.1999.809866

[66] S. Havre, B. Hetzler, and L. Nowell, "ThemeRiver: visualizing theme changes over time," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2000, pp. 115–123. doi: 10.1109/INFVIS.2000.885098

[67] H. Reijner, "The development of the horizon graph," in *Proceedings of the IEEE VIS Workshop on From Theory to Practice: Design, Vision and Visualization*. IEEE, 2008.

[68] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda, "Two-tone pseudo coloring: Compact visualization for one-dimensional data," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, Oct 2005, pp. 173–180. doi: 10.1109/INFVIS.2005.1532144

[69] W. Javed, B. McDonnel, and N. Elmqvist, "Graphical perception of multiple time series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 927–934, Nov 2010. doi: 10.1109/TVCG.2010.162

[70] M. Krstajic, E. Bertini, and D. Keim, "Cloudlines: Compact display of event episodes in multiple time-series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2432–2439, 2011. doi: 10.1109/TVCG.2011.179

[71] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski, "Visualizing time–oriented data–a systematic view," *Computers & Graphics*, vol. 31, no. 3, pp. 401–409, 2007. doi: 10.1016/j.cag.2007.01.030

[72] E. Garfield, "Citation indexes for science. a new dimension in documentation through association of ideas†," *International Journal of Epidemiology*, vol. 35, no. 5, p. 1123, 2006. doi: 10.1093/ije/dyl189

[73] E. Garfield, "The impact factor and using it correctly," *Der Unfallchirurg*, vol. 48, no. 2, p. 413, 1998.

[74] K. Börner and A. Scharnhorst, "Visual conceptualizations and models of science," *Journal of Informetrics*, vol. 3, no. 3, pp. 161 – 172, 2009, science of Science: Conceptualizations and Models of Science. doi: 10.1016/j.joi.2009.03.008

[75] S. Ghani, N. Elmqvist, and D. S. Ebert, "Multinode-explorer: A visual analytics framework for generating web-based multimodal graph visualizations," *Proc. of EuroVA'12*, pp. 67–71, 2012. doi: 10.2312/PE/EuroVAST/EuroVA12/067-071

[76] "Star metrics," https://www.starmetrics.nih.gov/, accessed: 2017-06-14.

[77] K. Madhavan, N. Elmqvist, M. Vorvoreanu, X. Chen, Y. Wong, H. Xian, Z. Dong, and A. Johri, "DIA2: Web-based cyberinfrastructure for visual analysis of funding portfolios," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1823–1832, Dec 2014. doi: 10.1109/TVCG.2014.2346747

[78] E. Halley, "An historical account of the trade winds, and monsoons, observable in the seas between and near the tropicks, with an attempt to assign the physical cause of the said winds," *Philosophical Transactions of the Royal Society of London*, vol. 16, no. 183, pp. 153–168, 1687. doi: 10.1098/rstl.1686.0026

[79] J. Snow, *On the Mode of Communication of Cholera*. John Churchill, 1855.

[80] A. H. Robinson, "The thematic maps of Charles Joseph Minard," 1967. doi: 10.1080/03085696708592302

[81] R. Burkhard and M. Meier, "Tube map: Evaluation of a visual metaphor for interfunctional communication of complex projects," in *Proceedings of the International Conference on Knowledge Management and Knowledge Technologies*, vol. 4. ACM, 2004, pp. 449–456.

[82] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "3D information visualization for time dependent data on maps," in *Proceedings of the International Conference on Information Visualisation*. IEEE, 2005, pp. 175–181. doi: 10.1109/IV.2005.3

[83] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploratory spatio-temporal visualization: an analytical review," *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 503–541, 2003. doi: 10.1016/S1045-926X(03)00046-6

[84] P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950. doi: 10.2307/2332142

[85] R. C. Geary, "The contiguity ratio and statistical mapping," *The Incorporated Statistician*, vol. 5, no. 3, pp. 115–146, 1954. doi: 10.2307/2986645

[86] A. Getis and J. K. Ord, "The analysis of spatial association by use of distance statistics," *Geographical Analysis*, vol. 24, no. 3, pp. 189–206, 1992. doi: 10.1111/j.1538-4632.1992.tb00261.x

[87] L. Anselin, I. Syabri, and Y. Kho, "Geoda: an introduction to spatial data analysis," *Geographical Analysis*, vol. 38, no. 1, pp. 5–22, 2006. doi: 10.1111/j.0016-7363.2005.00671.x

[88] G. L. Andrienko, N. Andrienko, D. Keim, A. M. MacEachren, and S. Wrobel, "Challenging problems of geospatial visual analytics," *Journal of Visual Languages & Computing*, vol. 22, no. 4, pp. 251–256, 2011. doi: 10.1016/j.jvlc.2011.04.001

[89] H. Djavaherpour, A. Mahdavi-Amiri, and F. F. Samavati, "Physical visualization of geospatial datasets," *IEEE Computer Graphics and Applications*, vol. 37, no. 3, pp. 61–69, 2017. doi: 10.1109/MCG.2017.38

[90] J. A. Wagner Filho, W. Stuerzlinger, and L. Nedel, "Evaluating an immersive space-time cube geovisualization for intuitive trajectory data exploration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 514–524, 2019. doi: 10.1109/TVCG.2019.2934415

[91] Q. Van Ho, P. Lundblad, T. Åström, and M. Jern, "A web-enabled visualization toolkit for geovisual analytics," *Information Visualization*, vol. 11, no. 1, pp. 22–42, 2012. doi: 10.1177/1473871611425870

[92] A. Godwin, Y. Wang, and J. T. Stasko, "TypoTweet Maps: Characterizing urban areas through typographic social media visualization." in *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*. Eurographics Association, 2017, pp. 25–29. doi: 10.2312/eurovisshort.20171128

[93] A. Godwin and J. T. Stasko, "Nodes, paths, and edges: using mental maps to augment crime data analysis in urban spaces," in *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*, Eurographics Association. Eurogrphics Association, 2017, pp. 19–23. doi: 0.2312/eurovisshort.20171127

[94] A. Slingsby and E. van Loon, "Exploratory visual analysis for animal movement ecology," *Computer Graphics Forum*, vol. 35, no. 3, pp. 471–480, 2016. doi: 10.1111/cgf.12923

[95] A. M. MacEachren, A. C. Robinson, A. Jaiswal, S. Pezanowski, A. Savelyev, J. Blanford, and P. Mitra, "Geo-twitter analytics: Applications in crisis management," in *Proceedings of the International Cartographic Conference*, 2011, pp. 3–8.

[96] J. K. Wright, "Problems in population mapping," *Notes on statistical mapping, with special reference to the mapping of population phenomenon*, pp. 1–18, 1938.

[97] D. Thom, H. Bosch, and T. Ertl, "Inverse document density: A smooth measure for location-dependent term irregularities," *Proceedings of the International Conference on Computational Linguistics*, pp. 2603–2618, 2012.

[98] L. Freeman, "The development of social network analysis," *A Study in the Sociology of Science*, vol. 1, p. 687, 2004. doi: 10.1016/j.socnet.2005.06.004

[99] P. Raghavan, "Social networks on the web and in the enterprise," in *Proceedings of the Asia-Pacific Conference on Web Intelligence*. Springer, 2001, pp. 58–60. doi: 10.1007/3-540-45490-X_6

[100] L. Garton, C. Haythornthwaite, and B. Wellman, "Studying online social networks," *Journal of Computer-Mediated Communication*, vol. 3, no. 1, p. JCMC313, 1997. doi: 10.1111/j.1083-6101.1997.tb00062.x

[101] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999. doi: 10.1145/324133.324140

[102] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, 2002. doi: 10.1177/016555150202800601

[103] A. Komarek, J. Pavlik, and V. Sobeslav, "Network visualization survey," in *Computational Collective Intelligence*. Springer, 2015, pp. 275–284. doi: 10.1007/978-3-319-24306-1_27

[104] N. Biggs, N. L. Biggs, and B. Norman, *Algebraic Graph Theory*. Cambridge University Press, 1993, vol. 67. doi: 10.1017/CBO9780511608704

[105] L. C. Freeman, "Visualizing social networks," *Journal of Social Structure*, vol. 1, no. 1, p. 4, 2000.

[106] M. Krzywinski, I. Birol, S. J. Jones, and M. A. Marra, "Hive plots—rational approach to visualizing networks," *Briefings in bioinformatics*, vol. 13, no. 5, pp. 627–644, 2012. doi: 10.1093/bib/bbr069

[107] M. Wattenberg, "Arc diagrams: Visualizing structure in strings," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2002, pp. 110–116. doi: 10.1109/INFVIS.2002.1173155

[108] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive sankey diagrams," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2005, pp. 233–240. doi: 10.1109/INFVIS.2005.1532152

[109] D. Bar-Natan, "On the Vassiliev knot invariants," *Topology*, vol. 34, no. 2, pp. 423–472, 1995. doi: 10.1016/0040-9383(95)93237-2

[110] M. Wattenberg, "Visual exploration of multivariate graphs," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 811–819. doi: 10.1145/1124772.1124891

[111] S. G. Eick and A. F. Karr, "Visual scalability," *Journal of Computational and Graphical Statistics*, vol. 11, no. 1, pp. 22–43, 2002. doi: doi.org/10.1198/106186002317375604

[112] S. Hadlak, H. Schumann, and H.-J. Schulz, "A survey of multi-faceted graph visualization," in *Proceedings of the Eurographics Conference on Visualization*. Eurographics Association, 2015, pp. 1–20. doi: 10.2312/eurovisstar.20151109

[113] G. J. Wills, "Nicheworks—interactive visualization of very large graphs," *Journal of Computational and Graphical Statistics*, vol. 8, no. 2, pp. 190–212, 1999. doi: 10.1080/10618600.1999.10474810

[114] T. Munzner, "H3: Laying out large directed graphs in 3d hyperbolic space," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 1997, pp. 2–10. doi: 10.1109/INFVIS.1997.636718

[115] J. Abello and F. Van Ham, "Matrix zoom: A visual interface to semi-external graphs," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2004, pp. 183–190. doi: 10.1109/INFVIS.2004.46

[116] M. Ghoniem, J.-D. Fekete, and P. Castagliola, "On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis," *Information Visualization*, vol. 4, no. 2, pp. 114–135, 2005. doi: 10.1057/palgrave.ivs.9500092

[117] N. Henry and J.-D. Fekete, "MatrixExplorer: a dual-representation system to explore social networks," *IEEE transactions on visualization and computer graphics*, vol. 12, no. 5, pp. 677–684, 2006. doi: 10.1109/TVCG.2006.160

[118] F. J. Newbery, "Edge concentration: A method for clustering directed graphs," in *Proceedings of the International Workshop on Software Configuration Management*. ACM, 1989, pp. 76–85. doi: 10.1145/72910.73350

[119] G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis, "Algorithms for drawing graphs: an annotated bibliography," *Computational Geometry*, vol. 4, no. 5, pp. 235–282, 1994. doi: 10.1016/0925-7721(94)00014-X

[120] J. G. Augustson and J. Minker, "An analysis of some graph theoretical cluster techniques," *Journal of the Association for Computing Machinery*, vol. 17, no. 4, pp. 571–588, 1970. doi: 10.1145/321607.321608

[121] D. H. Hutchens and V. R. Basili, "System structure analysis: Clustering with data bindings," *IEEE Transactions on Software Engineering*, no. 8, pp. 749–757, 1985. doi: 10.1109/TSE.1985.232524

[122] Y. Chung and A. Kusiak, "Grouping parts with a neural network," *Journal of Manufacturing Systems*, vol. 13, no. 4, pp. 262–275, 1994. doi: 10.1016/0278-6125(94)90034-5

[123] D. Auber, Y. Chiricota, F. Jourdan, and G. Melançon, "Multiscale visualization of small world networks," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2003, pp. 75–81. doi: 10.1109/INFVIS.2003.1249011

[124] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991. doi: 10.1002/spe.4380211102

[125] Y. Frishman and A. Tal, "Dynamic drawing of clustered graphs," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2004, pp. 191–198. doi: 10.1109/INFVIS.2004.18

[126] W. Chen, F. Guo, D. Han, J. Pan, X. Nie, J. Xia, and X. Zhang, "Structure-based suggestive exploration: a new approach for effective exploration of large networks," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 555–565, 2018. doi: 10.1109/TVCG.2018.2865139

[127] T. Major and R. C. Basole, "Graphicle: Exploring units, networks, and context in a blended visualization approach," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 576–585, 2018. doi: 10.1109/TVCG.2018.2865151

[128] P. C. Wong, P. Mackey, K. A. Cook, R. M. Rohrer, H. Foote, and M. A. Whiting, "A multi-level middle-out cross-zooming approach for large graph analytics," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009, pp. 147–154. doi: 10.1109/VAST.2009.5333880

[129] P. C. Wong, D. Haglin, D. Gillen, D. Chavarria, V. Castellana, C. Joslyn, A. Chappell, and S. Zhang, "A visual analytics paradigm enabling trillion-edge graph exploration," in *Proceedings of the IEEE Symposium on Large Data Analysis and Visualization*. IEEE, 2015, pp. 57–64. doi: 10.1109/LDAV.2015.7348072

[130] A. Perer and J. Sun, "Matrixflow: temporal network visual analytics to track symptom evolution during disease progression," in *Proceedings of the AMIA annual symposium*, vol. 2012. American Medical Informatics Association, 2012, p. 716.

[131] A. Srinivasan and J. Stasko, "Orko: Facilitating multimodal interaction for visual exploration and analysis of networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 511–521, 2017. doi: 10.1109/TVCG.2017.2745219

[132] S. Ko, I. Cho, S. Afzal, C. Yau, J. Chae, A. Malik, K. Beck, Y. Jang, W. Ribarsky, and D. S. Ebert, "A survey on visual analysis approaches for financial data," *Computer Graphics Forum*, vol. 35, no. 3, pp. 599–617, 2016. doi: 10.1111/cgf.12931

[133] I. Ajzen, T. C. Brown, and L. H. Rosenthal, "Information bias in contingent valuation: effects of personal relevance, quality of information, and motivational orientation," *Journal of Environmental Economics and Management*, vol. 30, no. 1, pp. 43–57, 1996. doi: 10.1006/jeem.1996.0004

[134] W. Aigner, S. Miksch, H. Schumann, and C. Tominski, "Survey of visualization techniques," in *Visualization of Time-Oriented Data*. Springer, 2011, pp. 147–254. doi: 10.1007/978-0-85729-079-3_7

[135] M. Novotný and H. Hauser, "Outlier-preserving focus+ context visualization in parallel coordinates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 893–900, 2006. doi: 10.1109/TVCG.2006.170

[136] N. Kong and M. Agrawala, "Graphical overlays: Using layered elements to aid chart reading," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2631–2638, 2012. doi: 10.1109/TVCG.2012.229

[137] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto, "WireVis: Visualization of categorical, time-varying data from financial transactions," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2007, pp. 155–162. doi: 10.1109/VAST.2007.4389009

[138] A. Malik, R. Maciejewski, N. Elmqvist, Y. Jang, D. S. Ebert, and W. Huang, "A correlative analysis process in a visual analytics environment," in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. IEEE, Oct 2012, pp. 33–42. doi: 10.1109/VAST.2012.6400491

[139] R. Amar, J. Eagan, and J. Stasko, "Low-level components of analytic activity in information visualization," in *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2005, pp. 111–117. doi: 10.1109/INFVIS.2005.1532136

[140] Y. Wang, F. Han, L. Zhu, O. Deussen, and B. Chen, "Line graph or scatter plot? automatic selection of methods for visualizing trends in time series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 2, pp. 1141–1154, 2018. doi: 10.1109/TVCG.2017.2653106

[141] P. Buono, C. Plaisant, A. Simeone, A. Aris, B. Shneiderman, G. Shmueli, and W. Jank, "Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting," in *Proceedings of the International Conference on Information Visualization*. IEEE, 2007, pp. 191–196. doi: 10.1109/IV.2007.101

[142] Y. Sun and M. G. Genton, "Functional boxplots," *Journal of Computational and Graphical Statistics*, vol. 20, no. 2, pp. 316–334, 2011. doi: 10.1198/jcgs.2011.09224

[143] W. G. Cochran and G. M. Cox, *Experimental Designs*. Wiley, 1950.

[144] B. Efron, "Bootstrap methods: another look at the jackknife," in *Breakthroughs in Statistics*. Springer, 1992, pp. 569–593. doi: 10.1007/978-1-4612-4380-9_41

[145] N. Schenker and J. F. Gentleman, "On judging the significance of differences by examining the overlap between confidence intervals," *The American Statistician*, vol. 55, no. 3, pp. 182–186, 2001. doi: 10.1198/000313001317097960

[146] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908. doi: 10.2307/2331554

[147] D. de Solla Price, "Editorial statements," *Scientometrics*, vol. 1, no. 1, pp. 3–8, 1978. doi: 10.1007/BF02016836

[148] L. Egghe and R. Rousseau, *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Elsevier Science Publishers, 1990. doi: 10.1086/602337

[149] C. Upson, T. Faulhaber, D. Kamins, D. Laidlaw, D. Schlegel, J. Vroom, R. Gurwitz, and A. Van Dam, "The application visualization system: A computational environment for scientific visualization," *IEEE Computer Graphics and Applications*, vol. 9, no. 4, pp. 30–42, 1989. doi: 10.1109/38.31462

[150] J. J. Thomas and K. A. Cook, "A visual analytics agenda," *IEEE Computer Graphics and Applications*, vol. 26, no. 1, pp. 10–13, 2006. doi: 10.1109/MCG.2006.5

[151] C. Larman and V. R. Basili, "Iterative and incremental developments. a brief history," *Computer*, vol. 36, no. 6, pp. 47–56, 2003. doi: 10.1109/MC.2003.1204375

[152] C. Heipke, "Crowdsourcing geospatial data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 550–557, 2010. doi: 10.1016/j.isprsjprs.2010.06.005

[153] G. Andrienko, N. Andrienko, P. Jankowski, D. Keim, M.-J. Kraak, A. MacEachren, and S. Wrobel, "Geovisual analytics for spatial decision support: Setting the research agenda," *International Journal of Geographical Information Science*, vol. 21, no. 8, pp. 839–857, 2007. doi: 10.1080/13658810701349011

[154] J. Kehrer and H. Hauser, "Visualization and visual analysis of multifaceted scientific data: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 3, pp. 495–513, 2013. doi: 10.1109/TVCG.2012.110

[155] M. Sarkar and M. H. Brown, "Graphical fisheye views of graphs," in *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. ACM, 1992, pp. 83–91. doi: 10.1145/142750.142763

[156] J. Zhao, M. Karimzadeh, L. S. Snyder, C. Surakitbanharn, Z. C. Qian, and D. S. Ebert, "Metricsvis: A visual analytics system for evaluating employee performance in public safety agencies," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1193–1203, 2019. doi: 10.1109/TVCG.2019.2934603

[157] H. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1946, vol. 43. doi: 10.1515/9781400883868-fm

[158] B. Shneiderman and C. Plaisant, "Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies," in *Proceedings of the AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*. ACM, 2006, pp. 1–7. doi: 10.1145/1168149.1168158

[159] S. A. Matei, M. G. Russell, and E. Bertino, *Transparency in Social Media*. Springer, 2015. doi: 10.1007/978-3-319-18552-1

[160] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton, "Can twitter replace newswire for breaking news?" in *Proceedings of the International AAAI Conference on Weblogs and Social Media*. AAAI, 2013.

[161] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the ACM International Conference on Web Search and Data Mining*. ACM, 2011, pp. 65–74. doi: 10.1145/1935826.1935845

[162] M. Hascoët and P. Dragicevic, "Interactive graph matching and visual comparison of graphs and clustered graphs," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 2012, pp. 522–529. doi: 10.1145/2254556.2254654

[163] M. Kumar, M. Hanumanthappa, and T. S. Kumar, "Crime investigation and criminal network analysis using archive call detail records," in *Proceedings of the International Conference on Advanced Computing*. IEEE, 2017, pp. 46–50. doi: 10.1109/ICoAC.2017.7951743

[164] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, pp. 2–es, 2007. doi: 10.1145/1217299.1217301

[165] J. S. Coleman, E. Katz, and H. Menzel, *Medical Innovation: A Diffusion Study*. Bobbs-Merrill Co, 1966. doi: 10.1002/bs.3830120608

[166] J. J. Gibson, "The theory of affordances," *Hilldale, USA*, vol. 1, no. 2, 1977.

[167] J. R. Maier and G. M. Fadel, "Affordance based design: a relational theory for design," *Research in Engineering Design*, vol. 20, no. 1, pp. 13–27, 2009. doi: 10.1007/s00163-008-0060-3

[168] R. Williams, *The Non-Designer's Design Book: Design and Typographic Principles for the Visual Novice*. Pearson Education, 2015.

[169] L. Wilkinson, *The Grammar of Graphics*. Springer Science & Business Media, 2013. doi: 10.1007/978-3-642-21551-3_13

[170] M. Chen and H. Jaenicke, "An information-theoretic framework for visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1206–1215, 2010. doi: 10.1109/TVCG.2010.132

[171] M. Chen and D. S. Ebert, "An ontological framework for supporting the design and evaluation of visual analytics systems," *Computer Graphics Forum*, vol. 38, no. 3, pp. 131–144, 2019. doi: 10.1111/cgf.13677

[172] M. Chen, M. Feixas, I. Viola, A. Bardera, H.-W. Shen, and M. Sbert, *Information Theory Tools for Visualization*. AK Peters/CRC Press, 2016. doi: 10.1201/9781315369228

[173] M. Chen, personal communication, 2020.

APPENDICES

A. SURVEY QUESTIONNAIRE: UNDERSTANDING THE KNOWLEDGE GAP BETWEEN DECISION MAKERS AND DATA ANALYSTS

Q1 **What is the average time analysts have to present their data on a decision to make to you?**

- Under 3 minutes: 0
- Under 5 minutes: 2
- Under 15 minutes: 0
- Less than half an hour: 2
- More than half an hour: 2

Q2 **What is the main difference in the way you and the analysts understand data?**

- "They have it in volume. I need it in highlights with the ability to ask for more"
- "I look at the data as how best to allocate my resources."
- "Analyst tend to focus on the manner in which data is captured where as end users tend to focus on the story the data is telling."
- "No difference. I see the data in the same light as the analysts."

Q3 **How often does the data presented change your decision?**

- Almost Never: 0
- Sometimes: 5
- Often: 1
- Almost Always: 0

Q4 **How often does the analyst include the raw statistics?**

- Almost Never: 2

- Sometimes: 2
- Often: 2
- Almost Always: 0

Q5 **How often do you wish to see the raw statistics?**

- Almost Never: 0
- Sometimes: 3
- Often: 1
- Almost Always: 2

Q6 **How much of an impact does seeing the raw statics make in understanding the data / making the decision?**

- Very Low: 0
- Some: 4
- Significant: 2
- Very High: 0

Q7 **Have you ever had an experience where the presented information appeared to be biased toward a decision?**

- Never: 1
- Occasionally: 4
- Often: 0
- Almost Always: 1

Q8 **Is there any additional information you think would be helpful for us to know in order to reduce the gap in transferring knowledge from the analysts to the decision makers?**

- "Understanding that the data is being captured consistently and is accurate is more important that interpretation."
- "From my perspective the data validates the decision to plan an operation, board vessels in a particular area."

- “I appreciate the opportunity to comment. I often brief senior level decision makers and do have have the luxury to display large amounts of raw statistics. I need to provide visual displays of information that convey my point. Depending on the point of view, we often try to provide our senior leader decision makers with preferred courses of action based on the visualized information. I align this with providing biased decisions, but if we convey the data appropriately, it should align with our recommendation.”

B. INTERVIEW QUESTIONNAIRE

- Traditionally, with this kind of dataset, how have you collected the information to make data-driven decisions?
- How might having this tool change the process and how you think about your decisions?
- Would any aspects of this system benefit you in making a data-driven decision?
- What improvements would you like to see to this work?
- What features does the current system have that you have not seen elsewhere?
- What information would you look for using this system? Or what would you want to look for in this dataset?

VITA

VITA

Calvin Yau is a Ph.D. student in the School of Electrical and Computer Engineering at Purdue University in West Lafayette, IN, USA. His research interests include visual analytics, information visualization, and human-computer interaction. He received his bachelor's degree in Electrical Engineering from University of Washington in Seattle, WA, USA.