VAST 2021 Mini Challenge 1: The Kronos Incident

Sai Krishna Chilvery
Computer Science
Arizona State University
Tempe Arizona USA
schilver@asu.edu

Dhanush Kukatla
Computer Science
Arizona State University
Tempe Arizona USA
dkukatla@asu.com

Hasya Udayan Patel
Computer Science
Arizona State University
Tempe Arizona USA
hpate131@asu.edu

Kavya Jignesh Parikh
Computer Science
Arizona State University
Tempe Arizona USA
kparik10@asu.edu

Khushkumar Kantaria
Computer Science
Arizona State University
Tempe Arizona USA
kkantari@asu.edu

Vedanti Dantwala
Computer Science
Arizona State University
Tempe Arizona USA
vdantwal@asu.edu

INTRODUCTION

In the VAST 2021 Mini-Challenge, our team focused on uncovering the complex relationships and events leading to a troubling situation involving GAStech, POK, the APA, and the Government. We analyzed a rich dataset, including news reports, GAStech employee resumes, and internal company emails. Our investigation centered on three key areas: identifying primary and derivative news sources and their interconnections; detecting biases in these sources, particularly in portraying specific people, places, and events; and exploring the intricate relationships among the key players. We utilized various visual analytics tools, like bar charts for news coverage and sentiment analysis, pie charts for individual source sentiment, a network link model for mapping relationships, a steam graph for tracking sentiment changes over time, and network relationship and timeline graphs for a broader view of the articles' connections. Our goal was to provide a comprehensive understanding of the events and dynamics leading to the mysterious disappearances, employing advanced analytical methods and visual storytelling.

Visualization Design

1. Network link model (Innovative chart):

Introduction:

The Network link model is a force-directed graph that is a combination of the bubble chart and the network model. It incorporates the categorical values the bubbles represent and visualizes links/connections between entities by combining parts of a network graph and bubble chart. It is a hybrid visualization that uses the size of the nodes to depict the quantitative aspects in addition to connections /interactions/relationships between various entities/nodes.

Purpose:

The primary motive for creating a hybrid visualization of bubble chart and network model is to enable consumers to understand the relationships between entities and their quantitative attributes in a single visualization, along with aiding in grouping, clustering, and pattern recognition depending on attribute sizes and connections. It fills the need in the market for a unified depiction of quantitative qualities as well as relational structures, enabling a deeper comprehension of intricate datasets and assisting in data-driven decision-making across a range of industries like Social Media, Marketing, Transportation, Customer Relations, Communication, Healthcare, and many other industries.

Explanation:

A node in the visualization represents the individuals inside an organization. The significance of an individual inside an organization is indicated by their circle node's size. For instance, the CEO and President are the most important people in the company hence their node will be the largest of all the nodes. The communication frequency between two linked members of the organizations is displayed by the linkages between the nodes. Hovering the tooltip over an individual inside an organization will display a deeper look of that person. As a result, every organization will be able to see who is present, and the connections will emphasize the communication between them.

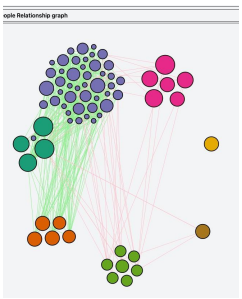

Fig 1: Network Link Chart

2. Word Cloud

Introduction:

A word cloud, also known as a tag cloud or text cloud, is a visual representation of text data where the importance of each word is shown with font size or color. This format is useful for quickly perceiving the most prominent terms and for getting an overview of large bodies of text. In a word cloud, frequently occurring words in the provided text are displayed in larger fonts or emphasized colors, while less frequent terms are displayed in smaller fonts.

Purpose:

The integrated word cloud in the VAST 2021 Mini-Challenge improves user involvement and provides deeper analysis inside the circular network graph. When links are clicked, multipurpose word clouds are displayed. First, they provide a quick comprehension of main topics by highlighting important themes in related articles. Word clouds also help identify potential biases based on phrase frequency and content focus. Additionally, users can compare word clouds for various links, providing insights into news providers' coverage perspectives. Users are able to explore and interpret the media landscape as part of this interactive feature, which adds an engaging dimension to data analysis.

Explanation:

When a user clicks on a link representing connections between similar articles from various news sources, the word cloud is automatically generated. This process involves extracting and processing text data from the connected articles, where algorithms identify and filter out common 'stop words', focusing on the most relevant terms. The frequency of each word's occurrence dictates its size in the cloud, allowing users to instantly recognize the most prominent themes. Additionally, color coding is employed to represent different themes or sentiments associated with the terms, adding another dimension to the analysis. For selected chains of articles, the word cloud adapts to reflect the collective content, offering a comprehensive view of the overarching themes and narratives. This functionality not only provides a rapid understanding of key topics and potential media biases but also encourages a deeper, more engaged exploration of the dataset, making it an invaluable asset in unraveling the complex narratives and relationships presented in the challenge.

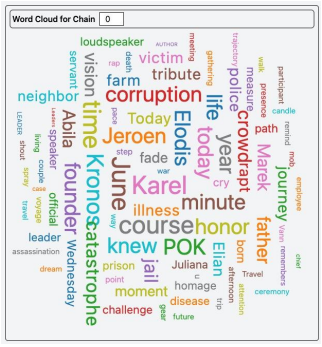

Fig 2: Word Cloud

3. Circular Network Chart

Introduction:

A circular network graph, or radial network diagram, is a visualization tool that arranges nodes (entities) around a circle's circumference, with connections (edges) shown as lines or arcs. This layout helps simplify complex networks by displaying relational dynamics between nodes. Nodes are strategically placed by criteria like alphabetically, by category, or based on network metrics. Interactive features in these graphs allow users to hover over or click on elements for more details, aiding in the exploration and understanding of network patterns and key connections.

Purpose:

The circular network graph in the VAST 2021 Mini-Challenge visualizes media coverage around GAStech and Kronos. Each news source is represented by an arc on a circle, with arc length indicating article volume. This design highlights active news outlets and their interconnections based on content similarity. Color-coding distinguishes between primary and derivative sources, providing insights into information origins. Interactive features enable users to explore article specifics and uncover themes and narratives.

Explanation:

The design of the circular network graph is carefully structured to enhance user understanding and interaction with the data. News sources are placed along the circle's circumference, each represented by an arc whose length is proportional to the number of articles published. This visual representation immediately informs users about which sources are most active in their coverage. The arcs are color-coded in two shades of blue, with a darker shade indicating primary sources of information and a lighter shade denoting derivative sources. This distinction is crucial for users to differentiate between original news reporting and content that is based on other sources.

Inside the circle, links connect arcs that represent news sources sharing similar articles. These links visually map out the relationships and content similarities between the sources, highlighting how information is shared or spread across the media landscape. The graph's interactivity comes into play when a user selects one of these links. Upon selection, a word cloud and a network timeline graph are displayed. The word cloud provides a quick summary of key terms and topics within the connected articles, while the network timeline graph offers a chronological view of how these topics have been covered over time. This level of interactivity not only makes the graph more engaging but also allows for a deeper exploration of the content, enabling users to uncover patterns and trends in the media coverage related to the events at GAStech and Kronos.

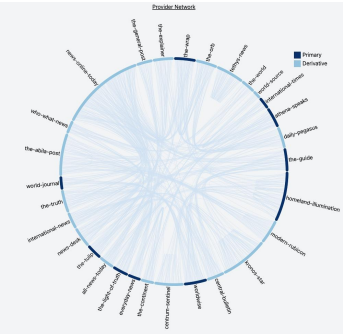

Fig 3: Circular Network Chart

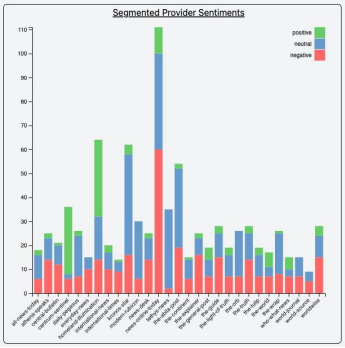

Fig 5: Sentiment Bar Chart

4. Bar Chart

Introduction:

A bar chart is a fundamental data visualization tool that displays discrete values in the form of bars, allowing for immediate comparison across categories. In sentiment analysis, this format becomes particularly valuable, offering a clear visual distinction between positive, neutral, and negative sentiments expressed in content from various publishers. Each bar's length or height correlates with the value it represents, enabling an intuitive understanding of the distribution of sentiments.

Purpose:

The primary objective of this bar chart is to provide a comparative analysis of sentiments across different publishers. It serves as an analytical tool to discern the prevalence of particular sentiment tones in published content, offering insights into the general attitude of each publisher's output. This analysis is crucial for stakeholders such as media analysts, advertisers, marketers, and readers who wish to understand sentiment biases in the media landscape.

Explanation:

This chart delineates sentiment counts for a selection of publishers, using a trio of colors—green for positive, blue for neutral, and red for negative sentiments. The x-axis enumerates publishers, while the y-axis quantifies the sentiment counts. The juxtaposition of coloured bars corresponding to each sentiment reveals the predominant emotional tone associated with each publisher at a glance. For instance, a taller green bar suggests a larger volume of positive content, whereas a dominant red bar indicates a greater count of negative sentiment. The ability to hover over bars to extract precise figures enhances the utility of the chart, making it a dynamic tool for sentiment tracking over time. Through this visualization, one can swiftly assess which publishers are more optimistic, neutral, or pessimistic in their communications, thus gaining valuable perspective on media bias and content tone.

5. Steam Graph

Introduction:

A streamgraph is a variant of a stacked area chart designed to display data changes over time around a central baseline, creating a flowing, stream-like visualization. Each layer in the graph represents a different data category, with their heights varying to reflect their values at different points in time. This structure allows for the visualization of continuous data patterns and facilitates easy comparison of changes across categories. The use of color helps differentiate these layers, enhancing visual clarity. Streamgraphs are particularly effective in handling complex multivariate time-series data, as they provide an intuitive way to discern trends and relationships.

Purpose:

In the VAST 2021 Mini-Challenge, the streamgraph effectively visualizes the changing sentiments towards key organizations like GAStech, APA, POK, and the Government from 1984 to 2012. Its primary function is to depict positive, negative, and neutral sentiments from news articles about these entities, aiding in identifying potential biases in media coverage. This visualization highlights shifts in sentiment over time, pinpointing moments of significant change that could mark key events or shifts in public perception. By showcasing trends and patterns in media sentiment, the streamgraph provides insights into possible reporting biases and offers a comprehensive view of how these organizations have been portrayed in the media over nearly three decades.

Explanation:

The streamgraph has 4 categories denoting the 4 organizations (GASTech, APA, POK and Government) which are represented using different colors. The x axis represents years ranging from 1984-2012 and the y-axis represents sentiments from the news articles provided by the sources associated with the organizations over a period of time.

Explanation:

Each node in the graph represents a different news article source.

The edges or connections between nodes signify the similarity or relationship between these sources. These could be based on shared content, referenced information, or similar sentiment trends. The graph's horizontal axis represents the timeline, showing the chronological progression of events and articles.

The vertical axis indicates the source of the news articles. The color of each node reflects the sentiment of the articles from that source, categorized as positive, negative, or neutral. Nodes are connected with curves to depict the relationship or similarity between different news sources.

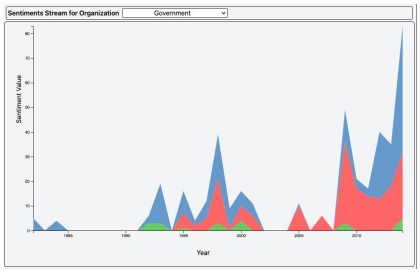

Fig 4: Network Timeline Graph

6. Network Timeline Graph

Introduction:

The Network Timeline Graph is an innovative visual analytics tool designed to explore and understand complex relationships within a dataset, especially when these relationships evolve. It is particularly useful in scenarios where temporal dynamics play a critical role, such as in analyzing news articles over time in the VAST 2021 Mini-Challenge.

Purpose:

The Network Timeline Graph is designed primarily to unravel and visually represent complex temporal relationships within data. Its key purpose is to illustrate how connections between various entities, such as news article sources, evolve over time. This is adeptly achieved by displaying nodes that represent different sources and using edges to highlight the links or similarities between them. These connections not only map out interactions among news sources but also bring to light the underlying patterns in their reporting. Additionally, the graph employs color-coded nodes to facilitate sentiment analysis, allowing users to track and analyze trends in sentiments (positive, negative, neutral) associated with each news source as they develop over the timeline. This aspect is particularly useful in providing a more transparent and comprehensive understanding of how different events are reported and perceived by various media outlets. In summary, the Network Timeline Graph is an essential tool for dissecting and comprehending the dynamics of news reporting, especially in scenarios where changes over time and relational connections between entities are crucial.

Explanation:

Each node in the graph represents a different news article source. The edges or connections between nodes signify the similarity or relationship between these sources. These could be based on shared content, referenced information, or similar sentiment trends. The graph's horizontal axis represents the timeline, showing the chronological progression of events and articles.

The vertical axis indicates the source of the news articles. The color of each node reflects the sentiment of the articles from that source, categorized as positive, negative, or neutral. Nodes are connected with curves to depict the relationship or similarity between different news sources.

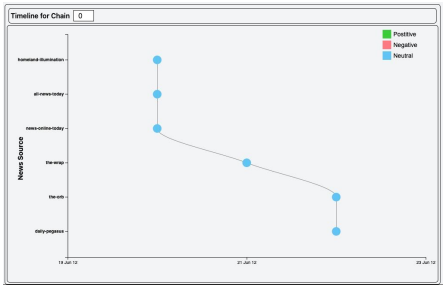

Fig 6: Timeline chart

Description of Vast MC Challenge

In the roughly twenty years that Tethys-based GAStech has been operating a natural gas production site in the island country of Kronos, it has produced remarkable profits and developed strong relationships with the government of Kronos. However, GAStech has not been as successful in demonstrating environmental stewardship. In January 2014, the leaders of GAStech celebrated their new-found fortune due to the initial public offering of their very successful company. Amid this celebration, several employees of GAStech went missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearances. It is January 21, 2014, and as an expert in visual analytics, you have been tasked with helping people understand the complex relationships among people and organizations that may have contributed to these events.

Dataset Description

The dataset provided for the VAST 2021 Mini-Challenge 1 is a rich and diverse collection of data types, each offering unique insights into the scenario surrounding GAStech, the island country of Kronos, and the mysterious disappearance of several employees. The data given for VAST Challenge 2021 Mini Challenge 1 includes the following files:

Historical documents:

These documents help understand the background and evolution of the entities involved, like GAStech, POK, and the government of Kronos. They offer insights into past events, decisions, and changes shaping the current scenario. This contains two text documents: A 5-year report titled The History of Protector of

Kronos (POK) summarizes the history of the POK. It assesses the likely future of the group and the profile of dominant POK personalities. The 10-year document highlights the history behind the entities and events of the POK that can offer clues to the root causes of the current situation, including the disappearance of GAStech employees.

Resumes of GAStech Employees:

The dataset contains resumes of employees working at GAStech, particularly those who went missing. These documents provide information on the employees' educational backgrounds, skill sets, professional experiences, and potentially their networks within and outside GAStech.

News articles:

The articles provide information about the unfolding events, public perceptions, and various narratives surrounding the situation at GAStech and the island country of Kronos. The articles contain a subject indicating the focus of the reporting, the main body of the news articles, providing detailed reporting on events, announcements, opinions, and analyses, and a published date, which is essential for understanding the timeline of events and media response and source details, i.e., information about the news provider, which can be critical for assessing the credibility and potential biases of the information.

Email headers:

A collection of email headers from internal communications within GAStech, spanning a critical two-week period. The headers include information like sender, recipient, date, and time of the emails.

Describe how to use your system to answer the MCs questions

1. Characterize the news data sources provided. Which are primary sources and which are derivative sources? What are the relationships between the primary and derivative sources?

In the VAST 2021 Mini-Challenge, the circular network chart is a crucial tool for characterizing news data sources, effectively distinguishing between primary and derivative sources, and elucidating their relationships. This categorization is ingeniously based on how often a news source initiates a chain of citations or references within the dataset. In the visualization, news sources are represented along the circle's circumference, each depicted as an arc. The key differentiator in categorizing these sources hinges on the frequency with which an article from a particular source starts a chain of references or citations. In this visualization, news sources are represented as arcs along the circumference of a circle, with primary sources - those that provide firsthand reports, direct statements, or original research - marked in one color, and derivative sources - which analyze, interpret, or cite primary information - in another. This color-coding, along with the strategic placement of arcs, enables a quick identification of each source's nature. Connections between these arcs illustrate the relationships between the sources, with links showcasing how often derivative sources refer to or cite primary ones, revealing the flow and transformation of information. The thickness and style of these connections further indicate the strength or frequency of these relationships, shedding light on which primary sources are most influential or widely referenced. This setup also highlights clusters of interconnected sources, suggesting shared perspectives or similar information pools, while isolated arcs may indicate unique or independent reporting. Moreover, the chart's interactive features allow for a deeper exploration of individual sources, offering insights into their specific articles and the nature of their influence in the broader media landscape. Overall, the circular network chart provides a comprehensive and visually intuitive means to understand the complex dynamics within the news ecosystem, crucial for the challenge's analysis.

The VAST 2021 Mini-Challenge utilizes a circular network chart to analyze news data sources. This visualization distinguishes primary sources, which offer original content, from derivative sources, which cite or interpret the primary information. News sources are represented as arcs along the circle's circumference, color-coded to differentiate primary from derivative sources. Connections between arcs indicate citation relationships, with the connection's thickness and style showing the strength and frequency of these links. This helps identify influential primary sources and clusters of interconnected sources, indicating shared perspectives. Isolated arcs suggest unique reporting. The chart's interactive features enable deeper exploration of individual news sources, their articles, and their impact on the media landscape, offering a clear, intuitive understanding of the news ecosystem's complex dynamics.

2.Characterize any biases you identify in these news sources, with respect to their representation of specific people, places, and events.

The analysis of news sources spanning from 1984 to 2012 in the VAST 2021 Mini-Challenge offers deep insights into the portrayal of specific people, places, and events, particularly concerning GAStech, POK, and other relevant entities. This investigation into potential media biases begins with a comprehensive sentiment analysis of the news articles.

Initially, the textual data from these articles undergoes pre-processing to standardize and clean it for analysis. This pre-processed data is then analyzed using the cardiffnlp/twitter-roberta-base-sentiment-latest sentiment analysis model, which classifies each article into categories of positive, negative, and neutral sentiments. The results of this sentiment analysis are visually represented through a stacked bar chart. This chart lists different publishers on the x-axis, while the y-axis quantifies the sentiment counts. Each sentiment category is color-coded—red for negative, green for positive, and blue for neutral—allowing an immediate visual grasp of the sentiment landscape across various publishers. Interactive elements are added to the chart; hovering over each segment reveals the

specific value of the sentiment, facilitating a more detailed exploration of the data.

In addition to analyzing the sentiment of the articles themselves, the sentiment analysis is extended to the representation of specific organizations mentioned in the articles. Proper and common nouns are extracted from each article, from which references to key organizations—GAStech, POK, the APA, and the Government—are filtered out. Sentiment values associated with these organizations are then predicted. This aspect of the analysis is visually depicted using a stream graph. A dropdown menu enables the selection of an individual organization, and the stream graph dynamically adjusts to display layers corresponding to the positive, negative, and neutral sentiments associated with the selected organization over time.

3. Given the data sources provided, use visual analytics to identify potential official and unofficial relationships among GASTech, POK, the APA, and Government. Include both personal relationships and shared goals and objectives. Provide evidence for these relationships.

The network link graph is adeptly used to visualize the intricate web of relationships within and between key organizations such as GAStech, POK, APA, and the Government. This graph, which effectively combines elements of a network graph and a bubble chart, serves as a dynamic tool for mapping both official and unofficial relationships among individuals and organizations.

The dataset for this challenge includes email headers, revealing 'to' and 'from' email IDs, and resumes that provide insights into the roles of individuals within their respective organizations. The roles indicated in the resumes are instrumental in determining each person's significance within their organization, with this importance visually represented in the graph by the size of the nodes. Larger nodes signify individuals with more significant roles, such as organization leaders, whereas smaller nodes represent other members.

Utilizing the data from the email headers, the graph illustrates the communication frequency between individuals, shedding light on unofficial relationships. These are inferred based on the frequency and pattern of email exchanges between two people. Official relationships, on the other hand, are deduced from correlations found in news articles, providing a comprehensive view of the formal connections between entities.

In this interactive network link model, organizations are represented as larger nodes, while the members within these organizations are depicted as smaller nodes clustered around their respective organization's node. The size of each member node varies according to their role's importance, creating a hierarchical visualization that is both informative and intuitive.

The links between nodes are color-coded to distinguish between the types of relationships. Red links denote official relationships, such as formal communications or publicly acknowledged connections derived from news articles. In contrast, blue links represent unofficial relationships, inferred from the analysis of email communication patterns.

This interactive model allows users to explore the network in detail. Clicking on a node can reveal more information about the individual or organization it represents, such as their role, their frequency of communication with others, and the nature of their relationships. This level of interactivity not only enhances the user's understanding of the network but also allows for a more engaged and exploratory analysis of the relationships within and across these organizations.

Discussion - Lesson learnt and Challenges

Throughout our engagement with the VAST 2021 Mini-Challenge 1, our team embarked on a comprehensive journey of technical skill enhancement, data processing, and storytelling through advanced data visualization techniques.

Data Processing with Python:

A substantial portion of our effort was dedicated to processing a large textual dataset, a task we navigated using Python. This experience was instrumental in enhancing our data wrangling skills, including cleaning, transforming, and preparing data for visualization. The use of Python allowed us to efficiently handle and interpret the complex dataset provided in the challenge.

Proficiency in Front-End Technologies:

The project demanded extensive use of HTML, CSS, and JavaScript, which are fundamental to any web-based application. In particular, our proficiency was significantly bolstered in utilizing D3.js for creating intricate and interactive data visualizations. The hands-on experience gained in manipulating SVGs and crafting dynamic, browser-based visuals was invaluable.

Storytelling and Communication through Visualizations:

One of the key learnings from this project was the art of storytelling through data visualization. We explored and implemented various types of visualizations such as network link model, timeline chart, radial chart, stream graph each offering unique insights into the dataset and answering the questions presented in the challenge. This process underscored the importance of selecting appropriate visualization techniques to effectively communicate complex data narratives.

Integration Challenges:

Each team member was tasked with developing different aspects of the visualization suite, leading to a diverse yet cohesive learning environment. However, integrating these individual components into a unified application presented challenges. We navigated issues related to code compatibility, data synchronization, and maintaining a consistent user experience. Overcoming these challenges was a significant learning curve, enhancing our capabilities in collaborative development and project integration.

Exploration of New Visualization Techniques:

We plan to explore and implement novel visualization techniques. This endeavor will broaden our understanding and skill set in the field of data visualization. By experimenting with untried methods, we aim to push the boundaries of our current capabilities and discover new ways to represent data more effectively and insightfully.

Conclusion:

The journey through the VAST 2021 Mini-Challenge 1 was a holistic learning experience for our team. It was an opportunity to solidify our technical skills in front-end development, data processing, nuances of visual storytelling and collaborative problem-solving.

REFERENCES

[1] https://d3-graph-gallery.com/network.html
[2] https://d3-graph-gallery.com/wordcloud.html
[3] https://d3-graph-gallery.com/chord.html