ClearWater: Interactive platform for bias investigation

Ana Luiza Nunes, Danilo Cardoso, Tomas Ferranti, Felipe Moreno-Vera, and Jorge Poco
{ana.lnunes, danilo.cardoso, tomas.ferranti, felipe.moreno, jorge.poco}@fgv.br
Fundação Getulio Vargas (FGV) - School of Applied Mathematics, Brazil

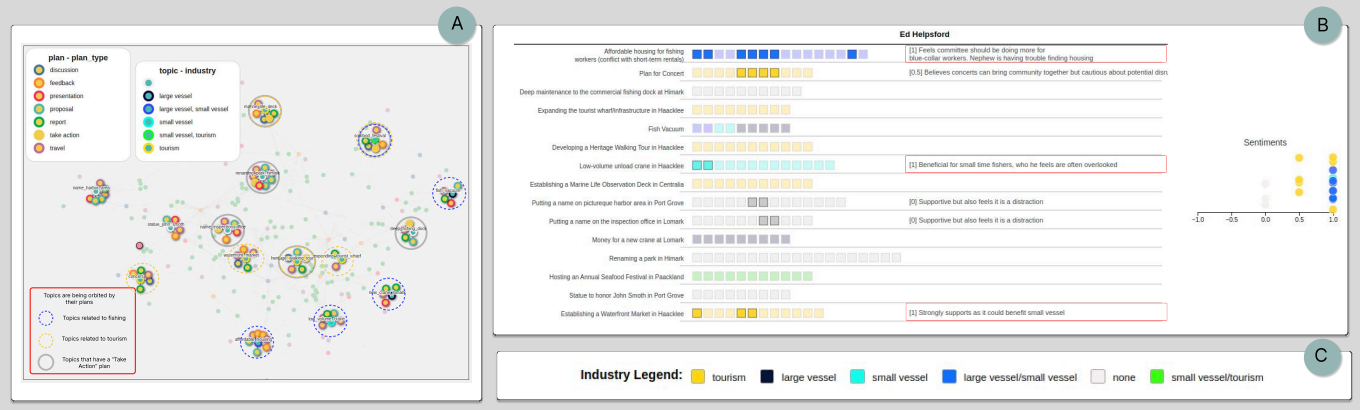

Figure 1: Visualization system overview: (A) Graph Visualization, (B) Members' Sentiments View, (C) Color Scheme, facilitates categorical comparison of topics.

Abstract

This paper introduces a visual analytics framework created for the VAST Challenge 2025, Mini-Challenge 2, aimed at investigating bias in the COOTEFOO committee. The system features three coordinated views: a graph for structural comparison, a member-centric view for individual contributions and sentiments, and a trip-oriented view for exploring temporal and spatial aspects. These views help identify both committee-level and sample-level biases in organizational datasets. The analysis uncovers a pro-tourism bias and selective omissions in the datasets, with the tool enabling transparent, data-driven evaluation of bias claims.

Index Terms: Visualization, Knowledge Graph, Bias detection

1 Introduction

The VAST Challenge 2025 Mini-Challenge 2 centers on journalist Edwina Moray's investigation into two independent datasets composed of meeting minutes and members' travel records: FILAH accuses the committee of favoring the tourism industry, while TROUT asserts a bias in favor of the fishing industry. Both datasets have allegations of bias against the COOTEFOO committee. Subsequently, Moray consolidated these materials, together with her own findings, into a unified dataset referred to as *Journalist*. Since broader contextual narratives may themselves reflect subjective biases, fine-grained representations of individual members' activities are essential for detecting personal alignments that may contribute to systematic committee-level bias. To address this challenge, we present a visual analytics tool called ClearWater (Fig 1) that enables the detection of bias both in the datasets provided by external organizations and in the committee's collective behavior.

2 Methodology

This challenge uses a dataset similar to those in previous studies [1, 2], involving the same country and a comparable goal of identifying illegal activities. However, it differs in the specific data and analytical approaches employed. Our methodology consists of two main stages: data preprocessing and visual analytics.

2.1 Data preprocessing

Topic classification: We aggregated all unique industry labels associated with members' sentiment nodes. This process yielded five distinct categories: *tourism*, *small vessels*, *large vessels*, *tourism/small vessels*, and *small vessels/large vessels*.

Spatial clustering: We clustered geographic locations using UMAP [3] and reconstructed map layout using those projections to enhance the interpretability of trip visualizations.

2.2 Visual Analytics System

Three views are employed in our approach:

Graph View: We created a two-layer interactive graph (Fig. 1.A), where users can select node types to display, with connected nodes automatically arranged in the lower layer. Nodes can be rearranged for better visibility. With three node types, the system uses an orbiting layout, bringing related nodes closer to the center. Each node's attributes can be highlighted by color. The *Highlight Dataset* feature adjusts node opacity to emphasize a specific subset while preserving the network context.

Sentiments and Reasons View: The sentiments and reasons view (Fig. 1.B) features a dual-panel interface. The left panel shows topics as a grid of colored squares, with color representing industry affiliation and opacity indicating participation. It also displays members' sentiments and justifications. The right panel features a scatterplot for sentiment analysis, with points representing sentiment scores from -1 (negative) to +1 (positive), colored by industry. The interface offers two modes: an overview of all members in a selected dataset or a comparative view with three panels for different datasets (Journalist, FILAH, TROUT).

Trips View: The trips view (Fig. 4) displays three horizontal rows per committee member, one for each dataset. Trips are shown as circles, with radius representing the number of stops and color indicating the year (red for 0040, blue for 2040). Trips are organized by weekday, helping users identify regularities and anomalies in travel patterns across datasets.

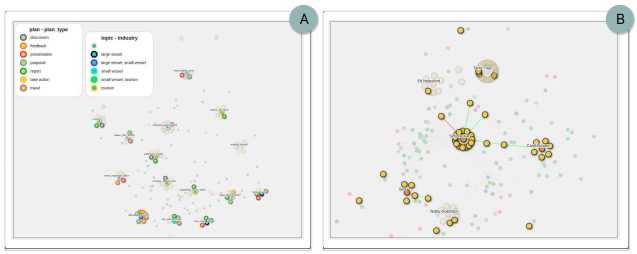

Figure 2: (A) Highlight TROUT over Journalist dataset, plan orbiting topic; (B) Highlight FILAH over Journalist dataset, plan orbiting person.

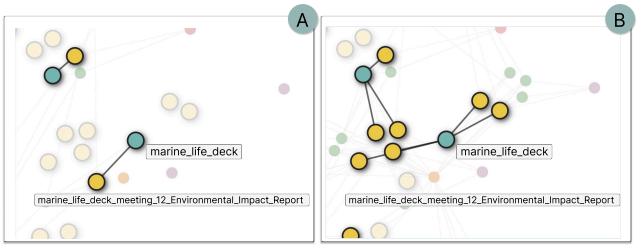

Figure 3: (A) Highlight TROUT over Journalist dataset, plan orbiting topic; (B) Highlight FILAH over Journalist dataset, plan orbiting topic.

Figure 4: Visualization of travel trajectories. The left panel displays three calendar-based heat maps (one per dataset), in which the radius of each circle encodes the number of stops within a trip and the color indicates the year (red for 0040, blue for 2040). The right panel depicts the spatial embedding of trip locations derived via UMAP, with dot colors representing different cities and red lines connecting the sequence of stops associated with the selected trip.

3 Results and Findings

This section presents the principal findings derived from our analysis, with particular emphasis on the identification of sample bias across datasets.

Bias of the Committee. Analysis of the consolidated Journalist dataset (Fig. 1.A) indicates a modest predominance of tourism-related topics (six versus five related to fishing), alongside ancillary agenda items—such as the designation of the Lomark inspection office—that, while industry-neutral, indirectly benefit tourism interests. Importantly, no fishing-related proposals progressed to the "Take Action" stage, suggesting that fishing initiatives were systematically constrained to discussion without resulting in actionable outcomes.

Voting patterns reinforce this trend. Carol, Simone, and Tante consistently supported tourism-related initiatives, while either abstaining from or opposing fishing-focused proposals. Simone's sole support for a fishing-related initiative—the installation of a small-vessel crane in Haacklee—was justified by its tourism benefits. In contrast, Teddy expressed exclusively pro-fishing and anti-tourism sentiments, while Chair Seal remained largely neutral. Ed's contributions (Fig. 1.B) tended to emphasize inclusivity, highlighting, for instance, the need to better support blue-collar workers. Nevertheless, his commentary often critiqued tourism-centered proposals for neglecting other constituencies. Taken together, these findings demonstrate a clear and consistent committee-level bias favoring tourism, with only isolated counterbalancing perspectives.

Bias in TROUT and FILAH. Overlaying TROUT data onto the Journalist dataset (Fig. 2.A) through opacity adjustments reveals a systematic omission of nearly all tourism-related discussions, especially "Take Action" plans. This exclusion creates a disproportionate emphasis on fishing activities and obscures tourism initiatives. Furthermore, all contributions from Carol, Simone, and Tante were absent from the TROUT subset, further reinforcing the distortion.

A parallel analysis of the FILAH dataset (Fig. 2.B) reveals the exclusion of all contributions from Ed, Teddy, and Tante. While FILAH's omissions appear to be more evenly distributed across industry categories than those of TROUT, the extent of removed material remains substantial.

A closer examination of "plan" nodes orbiting their associated "topic" nodes further underscores dataset discrepancies. TROUT retained exclusively the *Marine Life Deck Environmental Impact Report* (Fig.3.A), whereas FILAH preserved all other plans linked to the same topic, with the exception of this environmental impact report (Fig.3.B). These complementary omissions highlight divergent and selective filtering strategies by the two organizations.

Possible Meeting Locations. Fig. 4 suggests that most "0040" trips occurred on Tuesdays. Additionally, it is possible to identify sixteen places that were commonly visited by all members, which likely correspond to the 16 scheduled committee meetings. This inference is further supported by alignment with the known dates of meetings 13 and 15. Route analysis also identified two principal venues for committee gatherings: the Pacific Nature Bureau and the Jordan Administrative Center. Finally, anomalous timestamps indicating events occurring in the years 0040 and 2040 likely stem from data entry inconsistencies.

4 Conclusion

Our visual analytics framework reveals both committee-level bias and sample bias in the COOTEFOO case. By integrating structural, attitudinal, and spatiotemporal perspectives, the system makes omissions and systematic preferences directly visible. This enables transparent evaluation of bias allegations and provides a generalizable approach for supporting data-driven accountability in organizational contexts.

References

[1] D. Diaz, F. Moreno-Vera, J. Heredia, F. Venturim, and J. Poco. FishBiasLens: Integrating Large Language Models and Visual Analytics for Bias Detection . In *IEEE Visual Analytics Science and Technology VAST Challenge*. IEEE Computer Society, 2024. 1

[2] J. Heredia, F. Venturim, D. Diaz, F. Moreno-Vera, and J. Poco. Tracking Overfishing: Visual Analytics of Suspicious Behaviors in Commercial Fishing Vessels . In *IEEE Visual Analytics Science and Technology VAST Challenge*. IEEE Computer Society, 2024. 1

[3] L. McInnes, J. Healy, N. Saul, and L. Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. 1