Visual Analytics

Alessandro Carella

GitHub Project

Contents

1 Introduction

This report discusses the Mini-Challange 1 from the VAST challenge 2023 proposed by the teacher for this year Visual Analytics course.

One can read a summary of the scope of the challenge in the following:

Background FishEye International is dedicated to combating illegal, unreported, and unregulated (IUU) fishing and protecting marine species impacted by it. They gather online news articles about fishing, the marine industry, and international maritime trade. Using a natural language processing tool, FishEye extracts entities (people and businesses) and their relationships from these articles to create a knowledge graph. Analysts use this graph to investigate tips about possible IUU activities.

FishEye needs to quickly understand and explore the context around each tip to connect identified entities to illegal fishing activities and advance their investigations. They have included their own case studies in the database, making it possible for related organizations and companies to appear in the graph.

The challenge is to develop a visual analytics tool that helps analysts see, interact with, and understand the relevant context around a tip without being overwhelmed by unnecessary information. The display should be dynamic and interactive, emphasizing relevant contextual information to help identify companies potentially engaged in illegal fishing.

Tasks and Questions:

Use visual analytics to build a "contextualizer" that provides rich information for the entities listed below. Each entity was identified in a tip received by FishEye, but not all will end up with a substantive connection to illegal fishing. The visualizations you develop should allow analysts to extend the information they see as new entities are discovered.

Entities to investigate

The entities to investigate given by the challenge are are:

- Mar de la Vida OJSC,
- 979893388,
- Oceanfront Oasis Inc Carrie,
- 8327

Challange questions

The challenge questions are reported in Section 7 with the related answers.

2 State of the art

The approach that was taken for this project was to initially build a toolbox to explore the data visually since a clear pathway to solve the problem was not found. However, it does not mean that during the development useful references were not discovered.

Initially, an implementation with fixed coordinates for each node was developed based on the *Vega-Lite* airport connections which is one of the first examples founds during the initial phase of the project. This implementation was subsequently deleted due to the excessive amount of data of the challange.

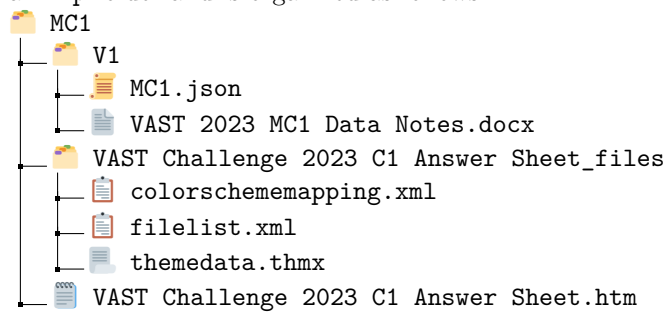
The idea of highlighting nodes sparked from this FlightConnections website. This website provides an overview of multiple flight routes throughout the whole world.

Regarding the implementation of the representation of the multilink between two nodes, as suggested by the teacher, the code inspired by this JSFiddle example. It is important to mention that the code was not used as it is, since it only covers the case where there are two types of links between two nodes, rather than the four in the case of the challenge data.

3 Description of data

3.1 Data structure

The VAST Challenge provides a dataset, which is accessible via this link. The dataset is contained in a `.zip` folder and is organized as follows:



The primary file of interest is `MC1.json`, which contains the data necessary for the challenge. The structure of the data is as follows:


```
{
  "directed": true,
  "multigraph": true,
  "graph": {},
  "nodes": [
      {
        "type": "company",
        "dataset": "MC1",
        "country": "Nalakond",
        "id": "Spanish Shrimp  Carriers"
      },
      ...
  ],
  "links": [
      {
        "type": "ownership",
        "weight": 0.90013963,
        "dataset": "MC1",
        "source": "Spanish Shrimp  Carriers",
        "target": 12744,
        "key": 0
      },
      ...
}
```


The data of primary interest includes the lists of nodes and links. From these, I selected the following attributes for my analysis:

- Nodes: type, country, and id.
- Links: type, weight, source, and target.

It was decided to remove the `dataset` attribute from both nodes and links, as its value was identical across all records given that all data originated from Mini Challenge 1. Additionally, the `key` attribute indicates the frequency of occurrences for the same source or target values, which was irrelevant to the representation.

To facilitate further analysis, two CSV files were created: `nodes.csv` and `edges.csv`, each containing the respective columns listed above.

3.1.1 Data preparation

Using the two datasets provided, an analysis of the information they contained was carried out. Some data cleaning techniques were implemented in both datasets to identify incorrect or irrelevant data and, consequently, reduce further analysis errors.

Nodes Several modifications were applied to the `nodes.csv` file, as outlined below:

- Removed nodes that did not appear as either a source or a target in the edges dataset;
- Identified that 1815 nodes were missing the type attribute;

- Identified that 6948 nodes were missing the country attribute;
- Checked for duplicates with `Pandas` method `df.duplicated().sum()` and discovered there were no duplicate rows in the nodes dataset;
- By iterating over the columns and applying `df[col].value_counts()` method, it was discovered that some IDs were repeated, as shown in Table 1.

Table 1: Duplicate IDs in the Node dataset.

Type	**Country**	**ID**
organization	None	18
vessel	None	18
organization	None	23
vessel	None	23
movement	None	38
organization	None	38
political_organization	None	621
vessel	None	621
political_organization	None	626
vessel	None	626
organization	None	77
vessel	None	77
event	None	90
organization	None	90

Initially, it appeared that "vessel" was the most common type among the duplicate entries. Based on this, the removal of the records where the type was "vessel", was initially considered assuming they were incorrect entries.

However, after visualizing the distribution of records for each type, represented in Figure 1, it became clear that "vessel" was not one of the most frequent types. As a result, it was decided to remove instances that were among the least common types in the dataset. For example, in cases like ID 18, of both type "organization" (with 987 instances in the dataset) and "vessel" (with 115 instances in the dataset), it was decided to keep the instance with the less frequent type ("vessel") in the new dataset.

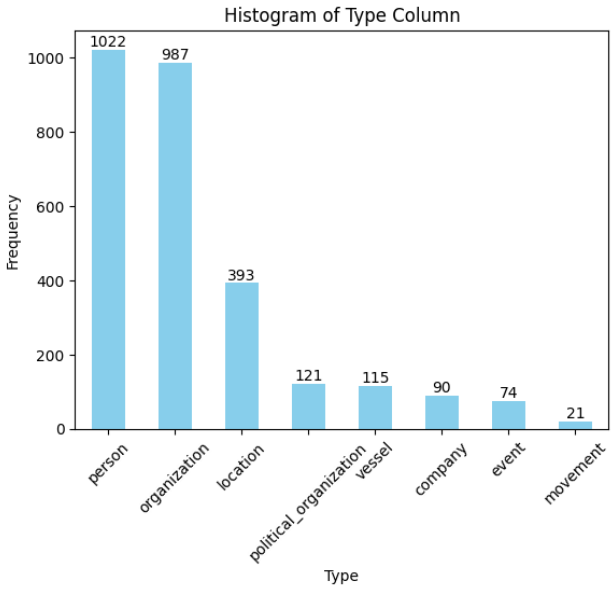


Figure 1: Classes of organizations plotted with frequency.

Edges It was found, from analyzing the data, that the majority of the weights of the links are above 0.8. In fact, out of the 11069 records, only 88 have a weight less than or equal to 0.8.

To uncover this, the distribution of weights was plotted using a bin size of 0.05 (noting that the weight variable ranges from 0 to 1). Figure 2 shows the resulting plot of the distribution.

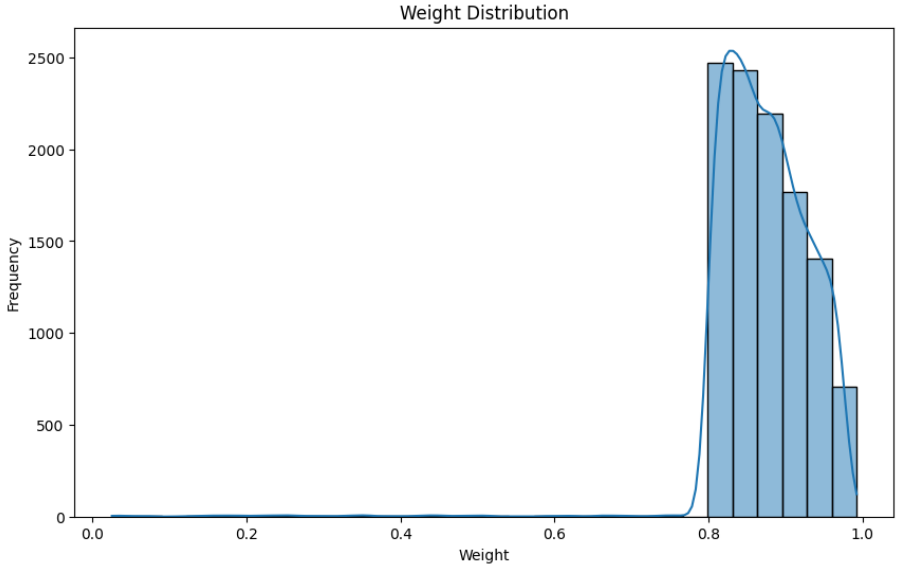


Figure 2: Weights distribution in edges dataset

In the proposed representation, the weight determines the thickness of the links between two entities. Given that low-weight links represent a very small portion of the dataset, their inclusion provides valuable context without significantly affecting the overall analysis, so it was decided to keep those links in the data.

Merged dataset After applying the previously mentioned modifications to both the edges and nodes datasets, they were merged, resulting in a combined dataset with the columns listed in Table 2.

Table 2: Merged dataset columns and missing values

Column Name	Number of missing Values
typeOfLink	0
weight	0
source	0
sourceType	2594
sourceCountry	8066
target	0
targetType	2769
targetCountry	7522

As discussed earlier, some nodes have missing values for the attributes type and country. Table 2 provides an overview of the columns and the corresponding number of missing values. To maintain the integrity of the dataset and ensure proper visualization, the missing values were replaced with the string "Unknown".

3.2 Presentation of the pattern or model to communicate

During the initial phase of data exploration of the project, a clear pattern that addressed the scope of the challenge was not identified. As a result, the scope of the project became to develop a user interface that would allow the user to find patterns through graphical data exploration.

As the UI took shape, it became increasingly evident that the key pattern to show to highlight illegal fishing involved clusters of entities that were highly interconnected, as well as entities with a high number of in and outgoing links, connected to these clusters.

4 Design choices

4.1 Colors

To make the graph more intuitive, colors were used to distinguish between different link types and node states. The challenge defines four types of links, each of which was assigned a distinct color: ownership, partnership, family relationship, and membership.

To accommodate various data exploration options, different colors were assigned to represent the five possible node states:

- Entity to investigate: Nodes representing the four specific entities listed for investigation in the challenge;
- Right clickable node: Nodes that, when right-clicked, open a context menu (described in subsection 5);
- Source and target: Nodes that display all their connections in the graph and act as both a source and a target;
- Source: Nodes that display all their connections and act only as a source, with no nodes pointing to them;
- Target: Nodes that display all their connections and act only as a target, with no outgoing connections.

With these definitions, a total of nine distinct colors were required for the representation. After testing various color palettes and tools suggested during the course, a qualitative color scheme from ColorBrewer was chosen with nine different colors. The palette can be seen in Table 3, and its application is illustrated in Figure 3.

Table 3: Final color palette.

	Name	Hex Code	Color Sample
Nodes	Entity to Investigate Node	#bc80bd	
	Clickable Node	#ffffb3	
	Source and Target Node	#b3de69	
	Source Node	#fb8072	
	Target Node	#8dd3c7	
Links	Ownership Links	#80b1d3	
	Partnership Links	#fdb462	
	Family Relationship Links	#b3de69	
	Membership Links	#fccde5	

Other colors that are not directly related to the graph and were arbitrarily chosen are in Table 4.

Table 4: Other colors.

Name	Hex Code	Color Sample
Link button disabled	#454545	
Number input plus button	#de3323	
Number input minus button	#2c1af4	
Opacity input bar	#395dc1	

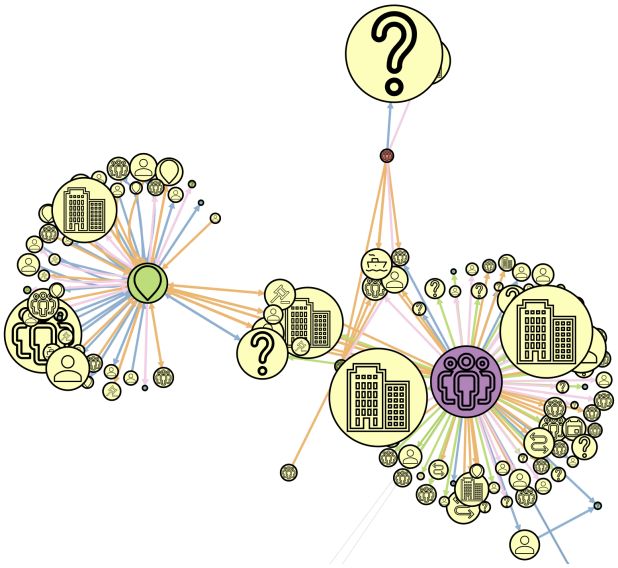


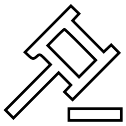
Figure 3: Chosen color palette applied

4.2 Shapes

4.2.1 Nodes

As shown in Figure 4 the primary shapes used in the graph are circles and links. To prevent overcrowding the visualization with too many colors icons, listed in Table 5, were used to differentiate between various node types. The SVGs of each icon were taken from SVG Repo.

Table 5: SVG icons for different types of nodes.

Type	Icon	Type	Icon
Unknown		Company	
Event		Location	
Movement		Organization	
Person		Political Organization	

Node sizes vary depending on the number of incoming and outgoing links, providing a visual cue for the amount of node's connections. Additionally, the node colors reflect the types of interactions users can perform with them, as previously discussed in subsection 4.1.

4.2.2 Links

The links in the graph include markers indicating their direction. If a link is bidirectional, markers appear at both ends. Additionally, when multiple types of links exist between two nodes, they are drawn at different angles to ensure all connections are visible simultaneously, as illustrated in Figure 4.

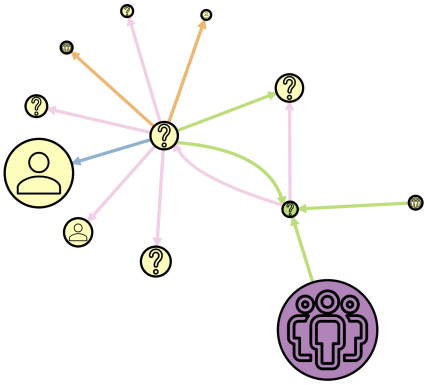


Figure 4: Multilink example.

The width of each link corresponds to its weight, with thicker lines representing stronger connections. If a link is bidirectional, its width is determined by the highest weight between the two directions. As discussed in subsection 3.1.1, the majority of weights in the dataset fall between 0.8 and 1, meaning the differences in thickness are often subtle, except in a few cases.

A dummy link was added between the entities under investigation. This link has a lower strength in the `D3.JS` simulation to prevent it from interfering with the user's navigation while still connecting the relevant entities.

5 Description of the interactions

Since the suspicious behavior of the entities was not identified during the initial data exploration, the process of building the visualization evolved into a "treasure hunt." The goal was to develop a comprehensive set of tools to enable effective visual inspection of the data.

As a result, the visualization provides users with multiple tools to navigate and explore the data, each designed to highlight potential patterns or behaviors. Most interactions modify the amount of data being displayed in the graph.

The one exception is the interaction triggered by clicking on a node. This action highlights the selected node and all nodes within a link distance of 1, as shown in Figure 5. More details about this interaction are provided in section 5.2.1.

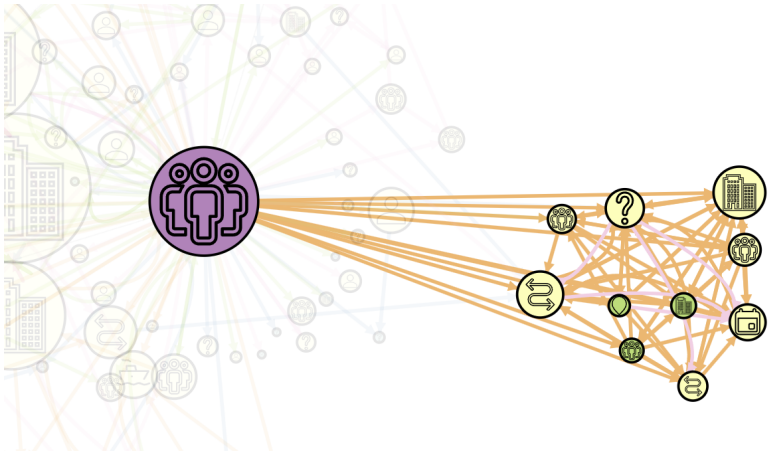


Figure 5: Node highlight example

Since it will be useful in the following sections the concept of Link distance n will now be introduced.

Link distance n Link distance n is defined as the distance from the node (or nodes) of origin as the amount of links needed to reach other nodes.

A practical example is shown in Figure 6, which represents a graph with mock nodes and connections.

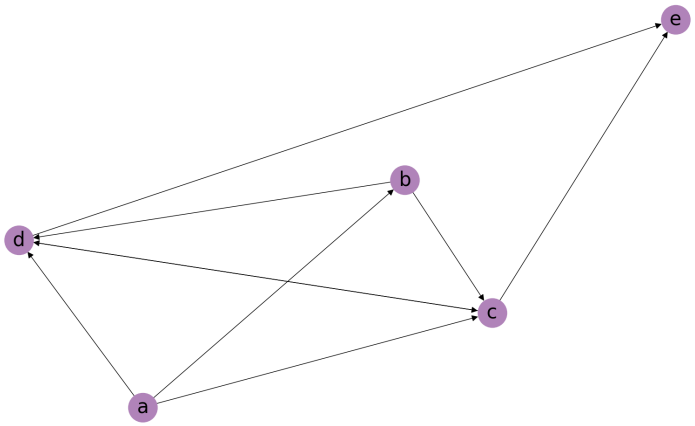


Figure 6: Simple graph to show the example graph in Link Distance Explanation

The following connections are represented on the graph:

- the node a connects to nodes b, c, d
- the node b connects to nodes c, d
- the node c connects to nodes d, e
- the node d connects to nodes c, e

Given these links, their representation on the `json` format used in the project would be:

```
[
    "a": {
        "1": ["b", "c", "d"],
        "2": ["e"]
    },
    "b": {
        "1": ["c", "d"],
        "2": ["e"]
    }
    "c": {
        "1": ["d", "e"],
        "2": []
    }
    "d": {
        "1": ["c", "e"],
        "2": []
    },
    "e": {
        "1": [],
        "2": []
    }
]
```

As illustrated, node a is connected to nodes b, c and d. Additionally, nodes c and d are connected to e. Therefore a has a link distance of 1 to b, c and d while it has a link distance of 2 to e since the connection between a and e is only via c and d.

5.1 Left Panel

The left panel serves as the core of the user controls, enabling users to interact with the visualization effectively. The following subsections detail the specific elements within this panel.

5.1.1 Investigate Number of Links Distance from Nodes to Investigate

As shown in Figure 7a this element of the panel includes two types of components: a dropdown select menu and three numeric inputs.

The select menu allows the user to choose which entity to investigate, with options include ”All” and the entities specified in the challenge: ”Mar de la Vida OJSC”, ”979893388”, ”Oceanfront Oasis Inc Carriers”, and ”8327”.

Once an option is selected, increasing one of the numeric inputs will reveal the nodes at link distance n from the chosen node. If a numeric input is adjusted without a selected value in the dropdown, the selection defaults to "All".

The three numeric inputs allow the user to investigate the target node from three perspectives: only as a source, only as a target, or as both. The default value for these inputs is -1, indicating that the corresponding input is disabled

These buttons are independent of one another. If a value is selected in one of the inputs, the others will be reset to -1. While a radio input for selecting the node perspective was considered, it was ultimately not implemented to avoid requiring multiple user inputs.

Despite being the last feature added to the visualization, this tool is positioned in the first available space as it effectively highlights unusual patterns in the entities' connections.

5.1.2 Choose Which Type of Links to Display by Clicking the Buttons

As shown in Figure 7b, there are four buttons available for the user to interact with. Each button corresponds to a specific type of link, disabling that type of link in the graph when clicked.

Once a button is pressed, the associated links will disappear from the graph, and the button will turn gray. Users can click the button again to show the links at any given time.

5.1.3 Choose the Opacity for Irrelevant Nodes and Links

When a node is clicked, the nodes at link distance 1 and their associated links will be highlighted. To highlight these connections, all other nodes and links will be displayed with lower opacity.

The opacity slider allows users to adjust this specific opacity, with a range from 0 to 100 to indicate a percentage, in increments of 10. In the code, the opacity value ranges from 0 to 1, so the selected percentage is divided by 100 when applied.

Users can set the opacity of irrelevant nodes to anywhere between 0% and 100%, with the default value set to 10%. The slider is disabled while a node is selected via a left click, as explained in section 5.2.1. If a user attempts to adjust the slider while a node is selected, a tooltip will appear with the message: "The slider is disabled when a node is selected." The slider can be seen in Figure 7c.

5.1.4 Specific Node Selection

Using the two select menus located at the bottom of the left panel, as shown in Figure 7d, users can select one of the possible options to view the corresponding node and its target nodes, or vice versa.

5.2 Graph

The graph displays the results based on the user's selections made in the left panel, described previously. It adjusts to each choice and offers the following additional interactions.

5.2.1 Left click

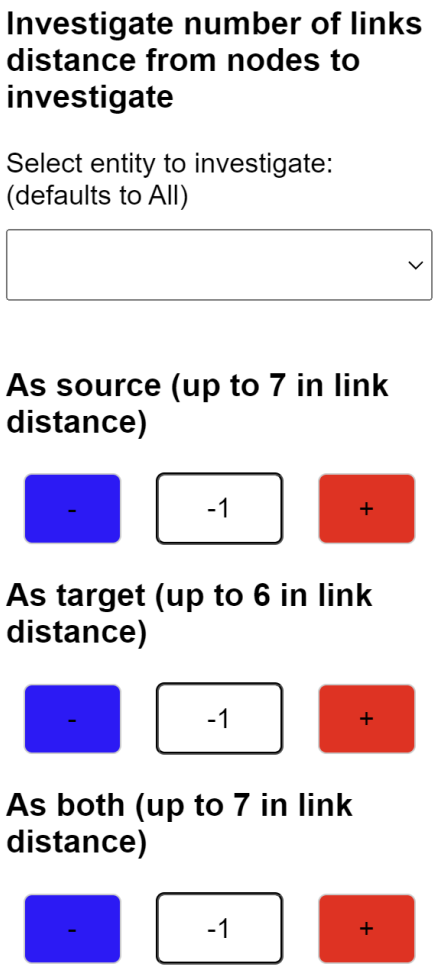
Clicking on a node will highlight that node along with its connections at link distance 1. An example of this interaction is illustrated in Figure 5.

To switch focus from one node to another, the user simply clicks on a different node, and its connections will be highlighted accordingly. To return to the default view, the user can click twice on the same node.

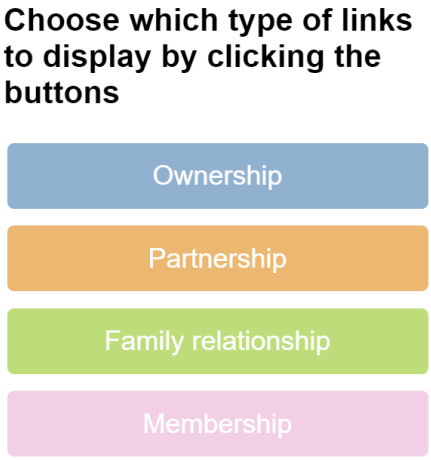
In previous versions of this feature, interested nodes were highlighted by changing their borders to red. This implementation was revised in favor of the current approach, as the opacity slider's functionality rendered the red highlight unnecessary unless the opacity slider is set to 100, which is ultimately a decision of the user.

5.2.2 Right click

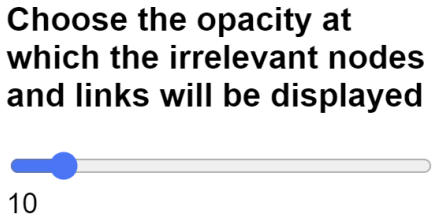
Right-clicking on a node opens a context menu that provides various options for the user. As shown in Figure 8, users can choose to expand the node as a source to reveal its targets, as a target to reveal its sources, or both.



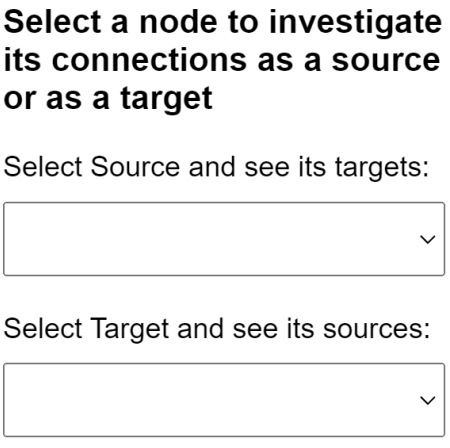
(a) Investigate number of links distance from nodes to investigate



(b) Link types buttons



(c) Opacity slider



(d) Selects to specify a source or a target

Figure 7: Left panel elements

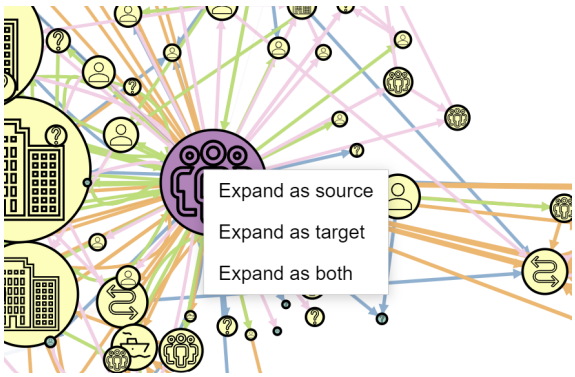


Figure 8: Context menu shown on the right click on a node event.

5.2.3 Hover

When the user hovers over a node, a tooltip appears, displaying contextual information related to that node. As illustrated in Figure 9, the tooltip contains the following data:

- The name of the node;
- The node type;
- The country of the node;
- The number of sources connected to the node;
- The number of targets connected to the node.

This tooltip is particularly relevant when combined with the right-click actions, as it provides precise information about the node before expansion. While the size of a node offers a visual hint of its potential size upon expansion, the tooltip delivers specific values for both interactions.

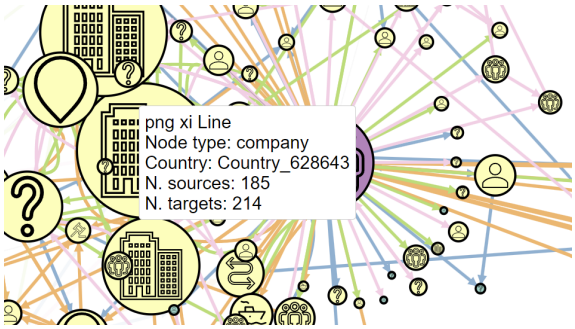


Figure 9: Tooltip example.

5.3 Map legend

In the top right corner of the graph, there is a button labeled ”Show Map Legend,” as shown in Figure 10. Clicking this button will display a right column, changing the button’s color and text to indicate ”Hide Map Legend.”

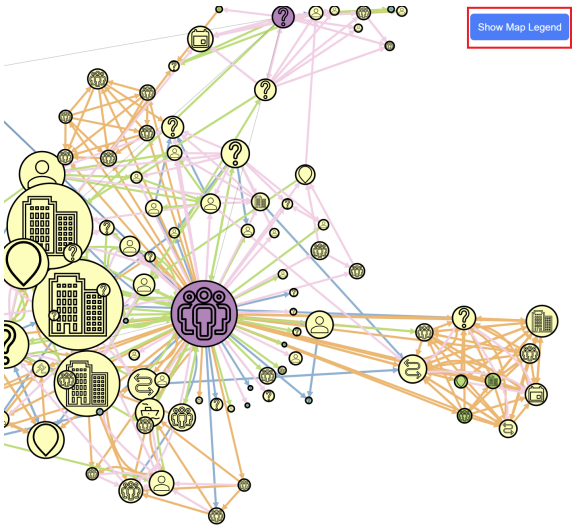


Figure 10: Map legend button.

In this column, users can read insights about the colors and symbols used in the graph. Additionally, it provides information regarding the left and right-click actions on the nodes, as well as details about the links present between all the entities being investigated, as specified in the challenge

6 Use Case Example
For an Analytical Task

This section presents an investigation of network connections between the main highlighted entities. By exploring the graph visualizations and interaction panels developed for this project, the aim is to uncover significant patterns and relationships between these entities, particularly focusing on the node *979893388*.

It is important to remark that the size of the nodes is dependent on the number of incoming and outgoing links. In the context of this and the following section (7), these nodes are often referred to as 'big' or 'large' to establish a direct visual connection with the images presented.

Using the left panel, specifically the subsection "Investigate Number of Links Distance from Nodes to Investigate", the select input was set to display all the entities. By increasing the number input value named "As Both" to 1, a view of all the links associated with the entities under investigation is displayed, including interactions where the node to investigate acts as both a source and a target.

As the user iterates through different types of links in the graph, it quickly becomes apparent that one of the nodes to investigate, *979893388*, is connected to several clusters. This becomes even more noticeable when only showing partnership-type links, as shown in Figure 11. Figure 12 highlights the distinct cluster formed by the big entities on the right of Figure 11. This cluster of major entities is well-connected to *979893388* through various intermediary nodes.

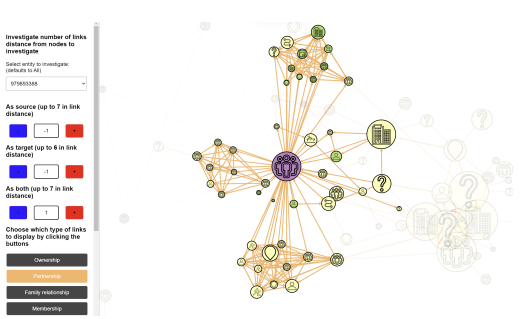


Figure 11: Clusters of partnership connections.

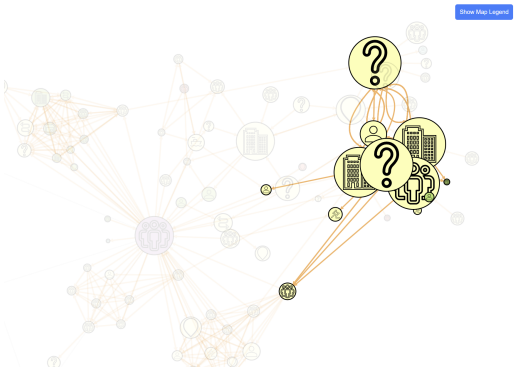


Figure 12: Cluster of big entities.

From the connections displayed, *Mar del Oeste Pic* is one of the main big nodes. Selecting

this node from the section named "Select a node to investigate its connections as a source or as a target" in the left panel, and expanding both its sources and targets shows its connections to other big nodes: png xi Line, Mar del Oeste Pic, Ocean Fisheries Llc, 903311212, SeaSpray Wave SRL Solutions. By expanding those big nodes, with all types of links enabled, a high degree of connections with *979893388* is revealed, as illustrated in Figure 14. However, an exception is the partnership link, which does not directly connect the major entities to *979893388*, as shown in Figure 14b.

Furthermore, *979893388* has a direct link with the following entities:

- Ownership:
  - *Mar del Oeste Pic*
  - *Ocean Fisheries Llc*
- Family relationship:
  - *png xi Line*
  - *Mar del Oeste Pic*
  - *Ocean Fisheries Llc*
  - *903311212*
  - *SeaSpray Wave SRL Solutions*
- Membership:
  - *png xi Line*
  - *Mar del Oeste Pic*
  - *903311212*
  - *SeaSpray Wave SRL Solutions*

Given these connections, the next step is to investigate whether any entity owns or shares other links with all five major entities listed above. Figure 13 shows the ownership connections of node *160*, which owns most of these entities, with the exception of *Mar del Oeste Pic*.

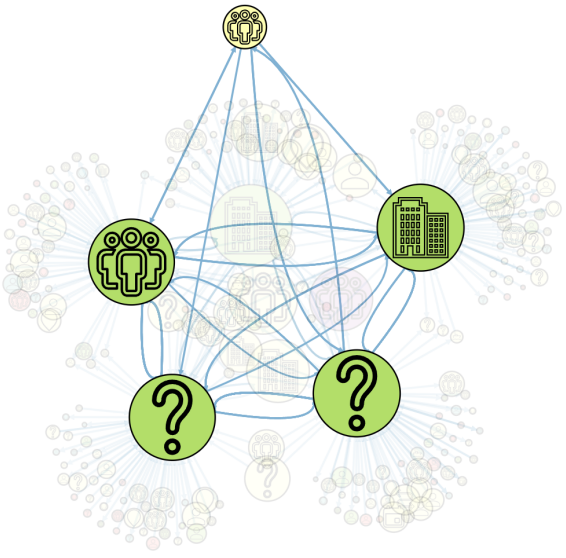
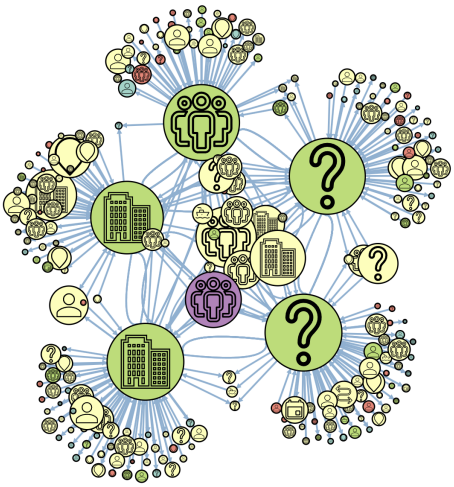


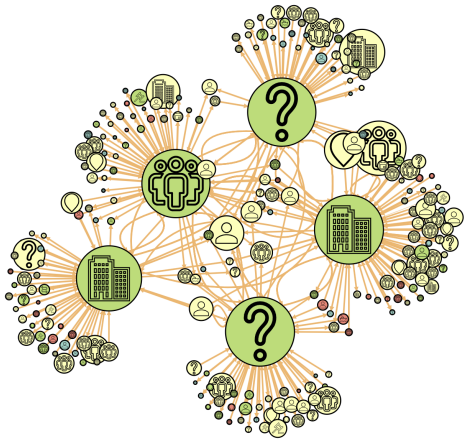
Figure 13: Entity *160* ownership connections.

When expanding the view of node *160* and visualizing its connections to other large nodes, it becomes apparent that these links are predominantly connected through ownership and membership, which means that the node has a broad ownership influence over all major entities except *Mar del Oeste Pic*. The connections are:

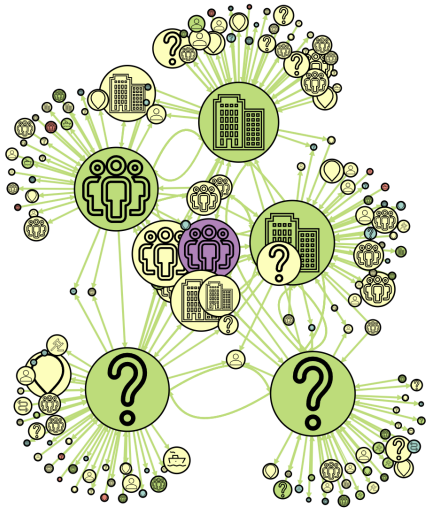
- Ownership:



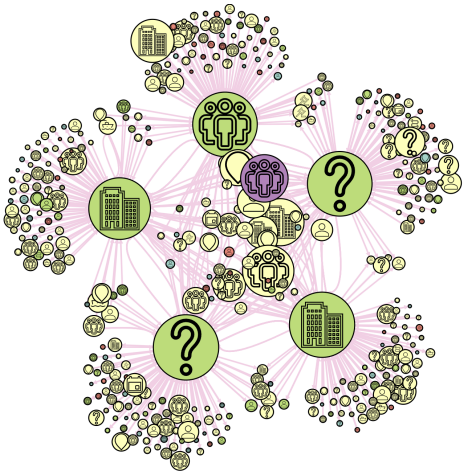
(a) Ownership.



(b) Partnership.



(c) Family relationship.



(d) Membership.

Figure 14: Multiple types of links between big entities.

- *Ocean Fisheries Llc* (owns and is owned)
- *903311212* (owns and is owned)
- *png xi Line* (owns)
- *SeaSpray Wave SRL Solutions* (owns)

- Membership:
  - *png xi Line* (from *160* towards *png Xi Line*)
  - *Mar del Oeste Pic* (from *160* towards *Mar del Oeste Pic*)

However, there are no links of type partnership or family relationship connecting *160* to any major entities, as illustrated in Figure 15.

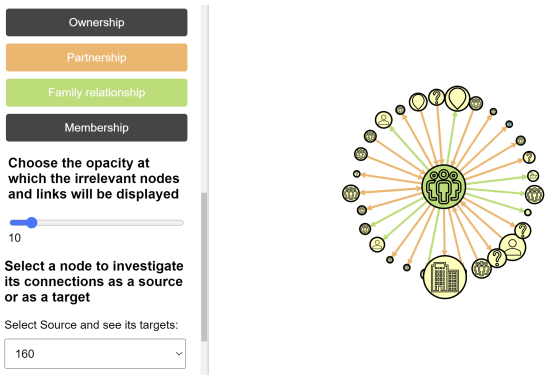


Figure 15: Partnership and family relationships of *160*

7 Answers to the Challenge Questions

Q1. *Use visual analytics to dynamically display and explore context around the suspected entities listed above. What did you learn about each one? Can you connect them to illegal fishing? Provide evidence for or against the case that each entity is involved in illegal fishing and use visual analytics to express confidence in your conclusions.*

Answer: By using the "Investigate Number of Links Distance from Nodes to Investigate" subsection of the left panel, one can explore the various connections between the entities under investigation. All nodes are connected to entity *97989338*, which itself has a lot of connected nodes, either directly or at a link distance 2. As a result, when the number input values are changed, the graph expands significantly as soon as one of the inputs is set to 2 or higher.

The ability to select and highlight a node to examine its connections at a link distance of 1 is crucial for detailed analysis. This feature allows users to gain insights into the immediate relationships of a specific node. To further enhance clarity, it may be beneficial to consider implementing an option to remove node *97989338*. This would enable a more focused examination of the connections involving other nodes without the distraction of the selected node's influence. However, a significant challenge arises from the fact that the lists of nodes at link distance n are precompiled rather than generated in real time, unlike other data in the program. This means that the feature implementation would not be effective.

Here is a brief analysis of all the remaining nodes in the list of the nodes to investigate:

- **Mar de la Vida OJSC**

  By setting the select input to the interested entity, and the number input named "As both" to 1, a cluster of entities connected to *Mar de la Vida OJSC* through membership-type connections is revealed, as highlighted in Figure 16.

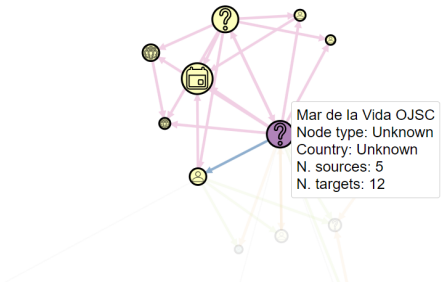


Figure 16: *Mar de la Vida OJSC* Membership cluster

All the nodes in this cluster are targets of *Mar de la Vida OJSC*. Switching the focus to the "As source" number input and increasing its value to 1, the visualization remained largely unchanged, confirming that the node primarily serves as a source while being linked mainly to smaller nodes. Increasing the "As source" input to a value of 2 revealed another cluster of entities at a link distance of two, connected by various types of relationships, represented in Figure 17, with node 33681303 acting as a middle-node.

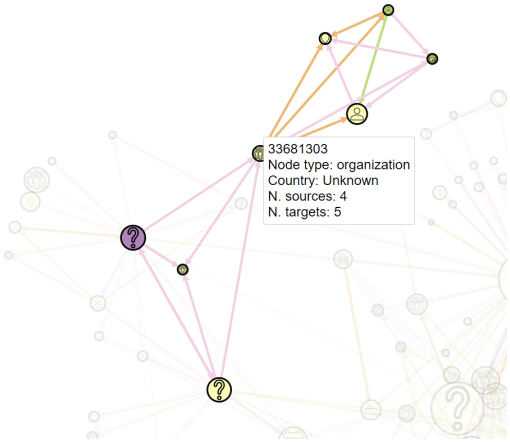
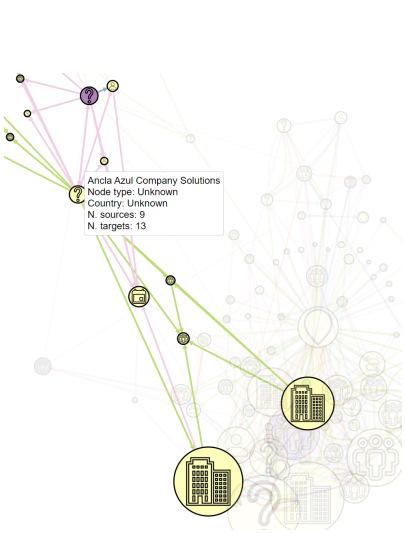
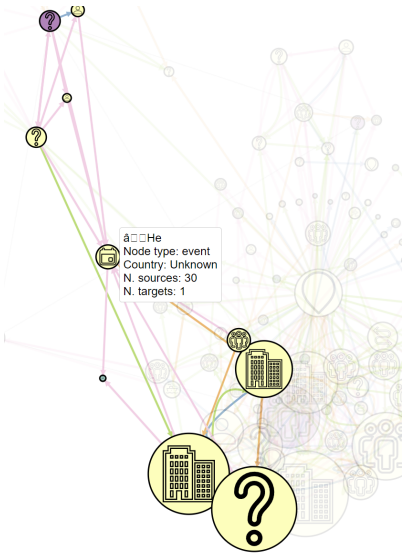


Figure 17: *Mar de la Vida OJSC* second small cluster

Two entities directly linked to *Mar de la Vida OJSC*: *Ancla Azul Company Solutions* and *â**He* are connected to larger entities discussed in Section 6, more specifically *Mar del Oeste Pic* and *SeaSpray Wave SRL Solutions*. This interconnection may suggest potential involvement in illegal fishing activities. Figure 18 highlights these connections.



(a) *Ancla Azul Company Solutions* Highlight



(b) *â**He* Highlight

Figure 18: Highlight of Mar de la Vida OJSC's connections to big entities

In summary, the network analysis made for entity *Mar de la Vida OJSC* shows significant interconnections with larger entities, which could indicate collaboration in illegal activities. However, more direct evidence of illegal practices by the node is needed to strengthen the conclusion.

- **Oceanfront Oasis Inc Carriers**

  This node presents a unique case, as it is only pointed to by other nodes without establishing any outgoing connections. When expanding the node using the "As both" number input and setting it to 1, no notable connections were revealed as shown in Figure 19.

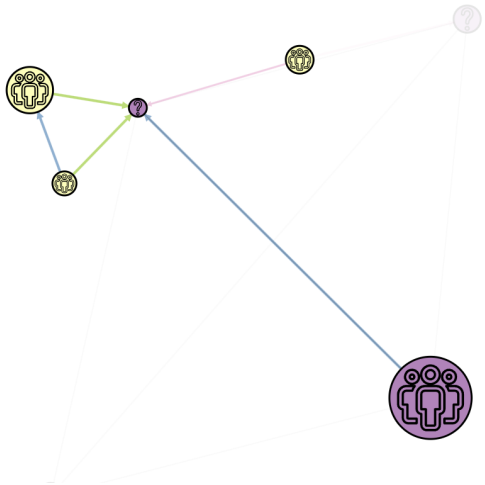


Figure 19: *Oceanfront Oasis Inc Carriers* link distance 1

The significant connection identified is with entity *979893388*, which owns *Oceanfront Oasis Inc Carriers* and was previously analyzed in Section 6. When exploring the connections at link distance 2, the graph expands dramatically due to the connections of *979893388*. Upon examining the nodes at link distance 1 from *Oceanfront Oasis Inc Carriers*, such as *FishEye International*, *2262*, and *8787*, it becomes evident that these nodes are linked to all major entities identified earlier, including *SeaSpray Wave SRL Solutions* and *Ocean Fisheries LLC*. This extensive network suggests a potential for involvement in illegal fishing. These connections are represented in Figure 20.

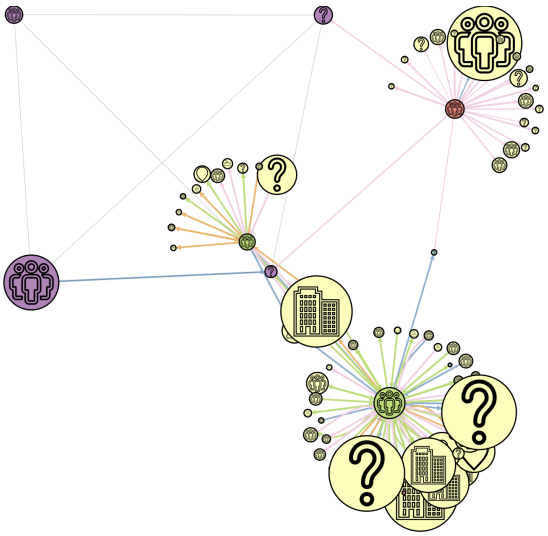


Figure 20: *Oceanfront Oasis Inc Carriers* link distance 1 expanded.

In summary, for entity *Oceanfront Oasis Inc Carriers*, the lack of outgoing connections and its ties to significant entities in the industry suggests that it is embedded in a network that may facilitate illegal fishing activities.

- **8327**
  For node 8327, using the same inputs as for the previous entities revealed two small clusters connected mainly through membership and ownership links at a link distance of 1, represented in Figure 21.

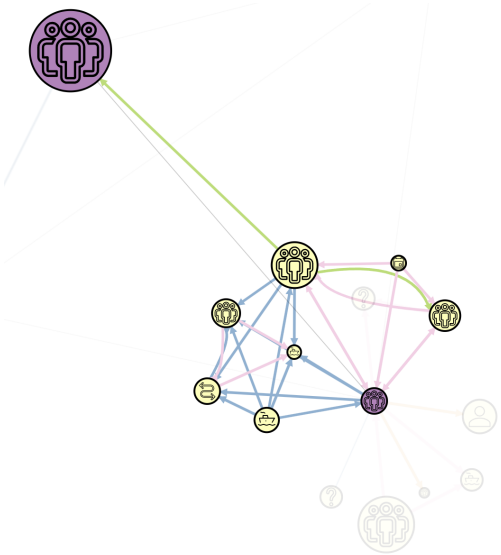


Figure 21: *8327*'s clusters at link distance 1.

Additionally, node 8327 connects at link distance 2 through the *Calvin Salas* node, using a partnership-type link to the cluster of large entities discussed in Section 6. Figure 22 illustrates these connections.

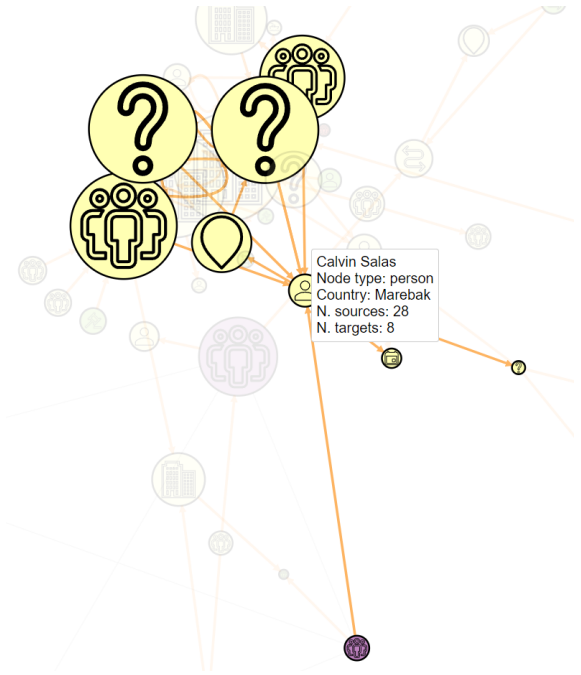


Figure 22: *8327*'s cluster at link distance 2.

It is also important to note that entity *8327* is also pointed to by entity *160*, investigated in Section 6. This layered connectivity indicates that *8327* may also have ties to illegal fishing activities.

Finally, for entity *8327*, while there are significant connections that suggest potential illegal activities, further evidence linking it to specific illegal fishing actions is necessary to make a more definitive conclusion.

Q2. *Use your visual analytics tool to compare and contrast what you learned about the suspect entities. Are there patterns that may indicate illegal activity? Use visual analytics to express confidence that a pattern exists and where uncertainty may be affecting this confidence.*

Answer: In the analysis using the visual analytics tool, a significant pattern associated with potential illegal activity was identified: the presence of clusters at a link distance of 1 or 2. These clusters suggest that certain entities may be interacting in a way that could indicate illicit behavior. Notably, the links between suspect entities and major entities, which have numerous connections, could imply illegal activities. However, it is important to recognize that these connections may also represent legitimate business interactions.

It is worth noting that the findings are based on patterns identified through the visual analytics interface developed rather than the prior data analysis, which did not yield valuable insights. This highlights the importance of visual exploration in uncovering potential irregularities. While the identified patterns suggest the possibility of illegal activity, further investigation is necessary to understand the nature of the connections and reduce uncertainty.

Q3. *What other companies should FishEye investigate for illegal fishing? Show how your visual analytics can be used to find entities that are worthy of further investigation.*

Answer: As previously remarked, the current interface of the visual analytics tool does not allow the user to explore entities beyond the four specified for investigation. This limitation arises from the fact that files related to these four entities are compiled prior to the run of the program, allowing for a focused analysis of their connections with other companies at link distance n.

To enhance the exploration of other potentially relevant entities, small modifications to the code can be made to include new files associated with different entities. However, this alternative route was not pursued for the following reasons:

- The compilation of the files for all the entities would have been too computationally expensive for the scope of the challenge

- Even with the compiled files, the time required for users to iterate through the data and identify significant results would be too long

Additionally, during the answer to the first question of the challenge, some clusters of entities not directly associated (at a link distance 2) with the four primary entities were identified.

To address the need for a more comprehensive analysis, an algorithm could be implemented to identify relative nodes based on the types of clusters discovered in the graph. While this approach would enhance the exploration of relationships and connections relevant to the challenge, it ultimately goes outside of the scope of the final part of the visual analytics project.

Q4. *Reflection: What was the most difficult aspect of working with this knowledge graph? Did you have the tools and resources you needed to complete the challenge? What additional resources would have helped you?*

Answer: As previously mentioned, the most challenging aspect of working with the knowledge graph was the initial data exploration phase. During this phase, it was difficult to identify patterns and relationships within the data prior to creating the visual representation. This lack of early insights made it harder to effectively direct the analysis and visualization that followed.

Additionally, this project marked my first experience using JavaScript and D3.js, which added to the complexity of the implementation. Without prior experience in these tools, I encountered a learning curve, especially when handling data without the support of an API. Relying solely on JavaScript for data handling limited the options and made the process less efficient, whereas an API would have facilitated faster data retrieval and greater flexibility.

Finally, using d3.js posed its own set of challenges, particularly implementing SVG icons within the circles of the graph, which proved to be complex and time-consuming, and adjusting the links to accurately represent multiple connections between nodes required applying angles based on the presence of other links. These were challenging aspects that demanded careful attention to detail.

Overall, while the project presented several challenges, identifying these difficulties highlights areas for improvement and potential future enhancements.