Introduction

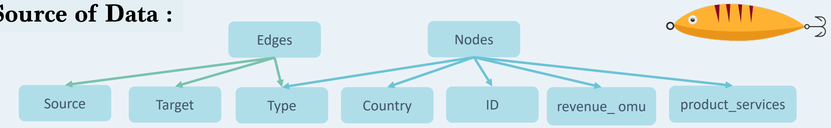
In this VAST Challenge 2023, FishEye International requires assistance in developing new visual analytics to identify anomalous companies that are involved in illegal fishing. In order to do so, we took up the challenge to develop a visual analytics process to find similar business groups.

An operational dashboard, network graph are created to allow user to visually analyze the international finance corporation's database on fishing related companies. The Shiny app will help user/ Fisheye International to solve these questions – *Which business group is unusual, what other type of business does they operate in, How similar the businesses are?*

Data Preparation & Methodology

The dataset was downloaded from VAST Challenge 2023's website. The dataset includes 2 files, namely the nodes and edges. Their relationship are shown below.

Source of Data :



Nodes are entities *(business type)* that need to be connected while Edges connects the nodes based on the defined layout. Likewise, Data Cleaning is done and only the aggregated/processed data were saved to ensure smooth deployment and lighter application load. *(File size <=30MB).* Additionally, data is processed R shiny application and imported back through *read_rds()* to avoid data overload.

Programming Language

The dataset was downloaded from VAST Challenge 2023's website and imported through the R package *'jsonlite'*. Topic Modelling uses package such as *'topicmodels'*, *'wordcloud'*, *'ldatuning'*. We have developed the application through various packages such as :

Shiny Application

Our application comprises of the following components:

1 Dashboard

Operative dashboard with **overview of Country , most identified words, and topics average revenue**

2 Network Analysis

Discover insights of business relationship patterns by looking at the community,

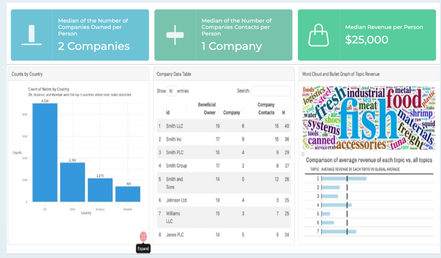
3 Topic Modelling

Identify various business group while measuring similarity and expressing confidence in the groupings

4 Deep-Dive Investigation

Based on various filters, user can conduct further investigations to visualized based on community, type, degree centrality

Dashboard



The dashboard is design to assist FishEye Intl in their investigation. It aims to identify company with anomalies and prompt users to walk through our website.

Country Count

Based on the number of nodes, we identified the top four countries.

Company Data Table

Represents the Top Company based on total count of Beneficial Owners, Company, and Company Contacts.

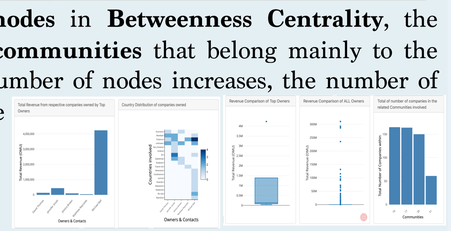
Word Cloud and Bullet Graph of Topic Revenue

Identify the top keyword extracted from the string of text in product & services. Words are grouped through topic modelling with bullet graph comparing the average revenue by topic.

Results

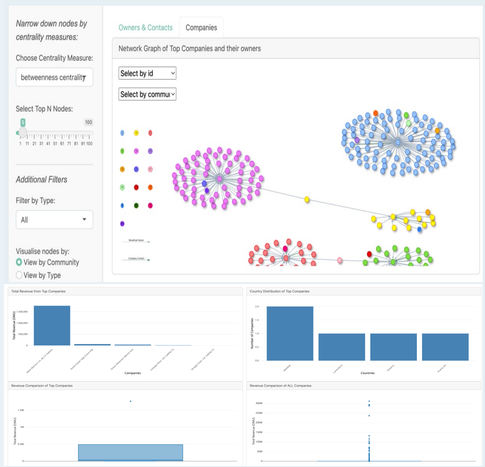
Network Analysis

Analysis on Company Ownerships and Contacts



With reference to the **top 5 nodes** in **Betweenness Centrality**, the projected network identified **9 communities** that belong mainly to the Beneficial Owners. When the number of nodes increases, the number of distinct community increases. We deep dive into the revenue and countries. From the first graph *(left)*, we identified that **Micheal Bell** have the highest revenue, and is involved in multiple countries *(5th row)* with concentration in Oceanus. Thereafter, we looked at the 2nd graph *(right)* to compare the revenue of the top 5 nodes against all owners. It is evident that the revenue of all owners are **not heavily dependent** on the top 5 nodes. Likewise, we note that the top 5 nodes belong to **4 different communities.**

Analysis on Top Companies and their owners

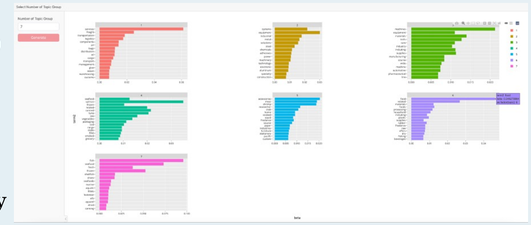


Similarly, the projected network reference to the **top 5 nodes of Companies** in **Betweenness Centrality.** The companies are well segregated into 5 different clusters. As the number of nodes increases, cluster will become less distinct. User will be able to observe the total revenue, the company's country *(enable graph select)* and compare the revenue against all companies.

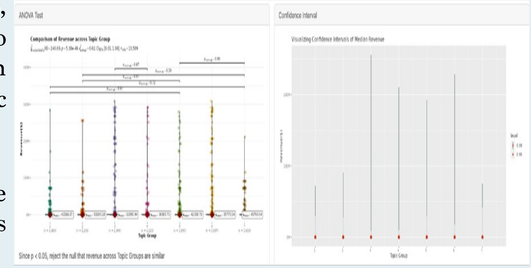
Topic Modelling

Analysis on Topic Groups

Based on the **Latent *Dirichlet Model (LDA)*** model, we identified the optimal number of topic groups to be 7.



To measure the similarity among the topic group, ANOVA test is used to test if the the median revenue across the topic group is the same.

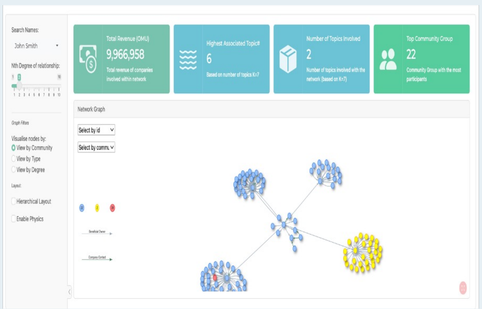


Since P-value (<0.05), we reject the null hypothesis that the median revenue is the same. For Pairwise- comparison, there are pairs with values >0.05. As such, we cannot reject the null hypothesis that there is no differences between the topic group and revenue. By visualizing the confidence interval, it is evident that topic 3 and 6 exhibits a higher confidence.

Deep-Dive Investigation

Analysis of ID by N-th degree of relationship

To aid FishEye Intl in further investigation, user will be able to search based on ID and choose the level of degree that they are interested in.



In this example, we search for **John Smith**, and to look at **2nd degree of relationship** with the inclusion of graph filters to view by community, type, and degree. After selecting the various options, the insights will bring about the total revenue, highest associated topic, number of topics involved, and the top community group. We note that that John Smith is in 2 topic groups but mostly associated with Topic 2. For better visualization, hierarchical layout is introduced.

Future Work

Due to size constraint on Github and R Shiny Application, constraints are put in place to limit the number of nodes (*100*) and number of topic groups (*20*). Despite so, we believe that by creating new visualization, we can enhance the analysis and elevate the interactive experience .

Moreover, FishEye could propose to retrieve/add time period from the international finance corporation's database. It would allow the team to conduct time-series analysis based on the its revenue.

KIM JUNSEOK | OH JIA WEN | THNG YAO WEN | Group 11