Vast Challenge 2023
Uncovering Illegal, Unreported, and Unregulated (IUU) through Visual Analytics

Chen Fangxian | Huynh Minh Phuong | Dabbie Neo Wen Min

Introduction: FishEye International aims to combat IUU fishing by utilizing a comprehensive database on fishing-related companies. They have observed that atypical company structures often indicate involvement in IUU activities. To leverage this resource, FishEye has transformed the database into a **knowledge graph**. However, traditional visualizations and analyses are insufficient due to the complex and extensive data. This project aims to develop an **innovative visual analytics approach to detect fishing business anomalies and accurately identify companies engaged in IUU fishing.**

1. Explore Ownership Network

1.1 Distribution of Revenue and Number of Owned Companies

Number of owned companies for each target is calculated by grouping the target, source and type in mc3_edge, and filtering type == 'Beneficial Owner'.

Left joined of the mc3_nodes and mc3_edges is performed to include revenue_omu information.

A **scatter plot** using *ggplot* displays count of companies on the x-axis and revenue on the y-axis, with tooltips enabled via *ggplotly*.

DT::datatable( ) function is utilized to generate a table with target, count_of_companies, and revenue information. *Crosstalk package with bscols* function is used to create the connected plot and data table.

Show 5 entries Search: michael Johnson

	target	count_of_companies	revenue
2	Michael Johnson	9	104221758

Showing 1 to 1 of 1 entries (filtered from 9,898 total entries) Previous 1 Next

1.2 Network Graph of Target ID

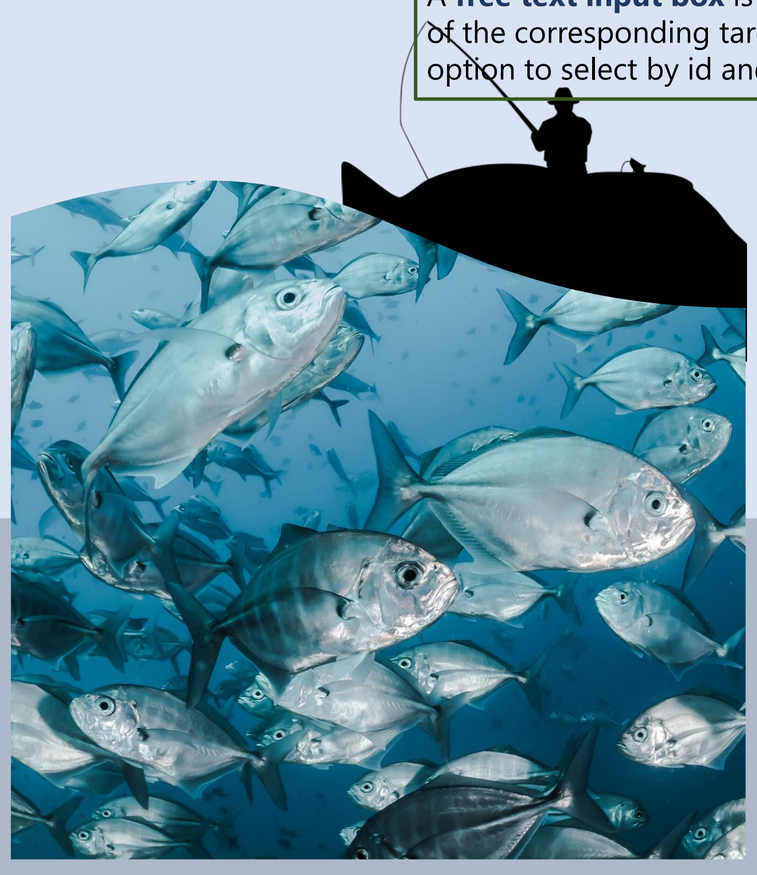
Unique targets and sources with their types are combined to get the nodes. It is then used to create *a tibble graph* along with the edges.

A group id is assigned to all connected nodes using the code *snippet 'mutate(group_id = components(mc3_graph)$membership)'*, which adds a new column named "group_id" to the node dataframe. The group_id is determined based on the membership of each node in the components of mc3_graph. The *components()* function returns a list of components in the graph, and the *$membership* part extracts the membership vector indicating which component each node belongs to.

Three functions are created: one to return a subgraph based on a selected group id, another to retrieve the group_id for a given target, and the last one to return the network graph of a target.

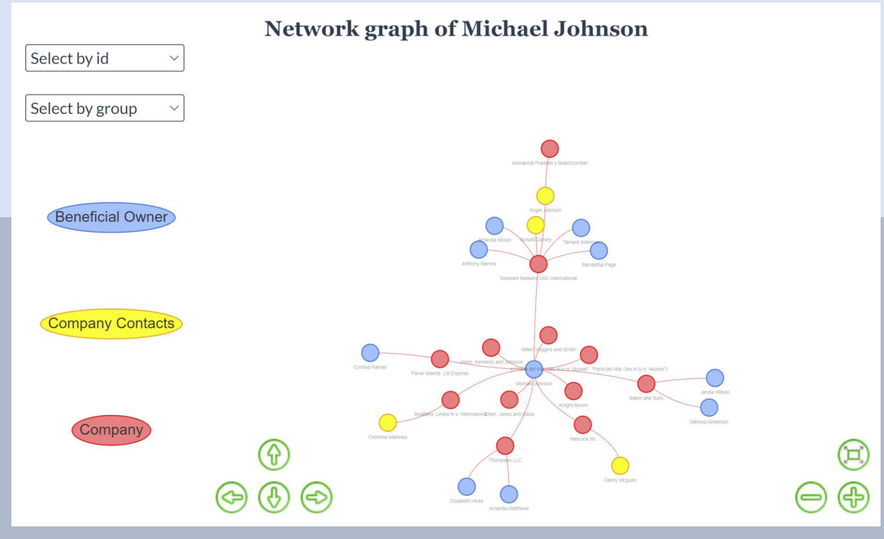
VisNetwork package is used to plot the interactive network graph.

A **free text input box** is created for users to enter Target Name and the network graph of the corresponding target name will be displayed automatically. Users are also have the option to select by id and group in the network plot itself.



Enter Target Name (*case-senstive):

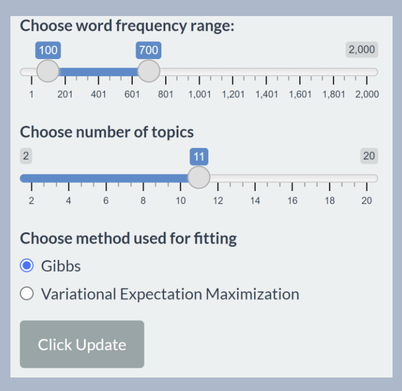
Michael Johnson



2. Explore Company Industries and Revenue

To provide interactivity, adjustable **sliders** are created for word frequency range and the number of topics and users can select the method to use for fitting for topic modelling.

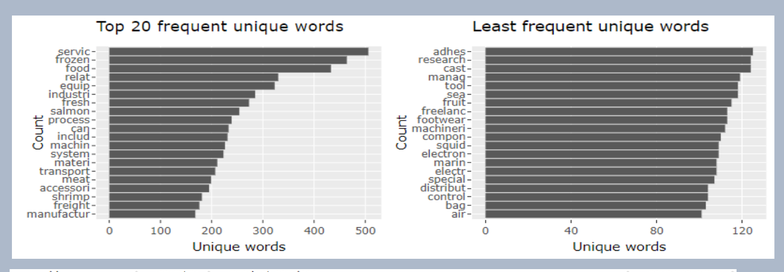
All charts and table in the different sub tabs will be updated based on your chosen parameters



2.1 Text Sensing with tidytext

The description of products and services for each company was tokenized. Stop words were removed and stemming was applied. The high and low frequency words are also removed for noise reduction and better computational efficiency.

Bar charts were generated to display the top 20 most and least frequent unique words. Additionally, a **dynamic table** was created to provide insights on the products offered by different companies.

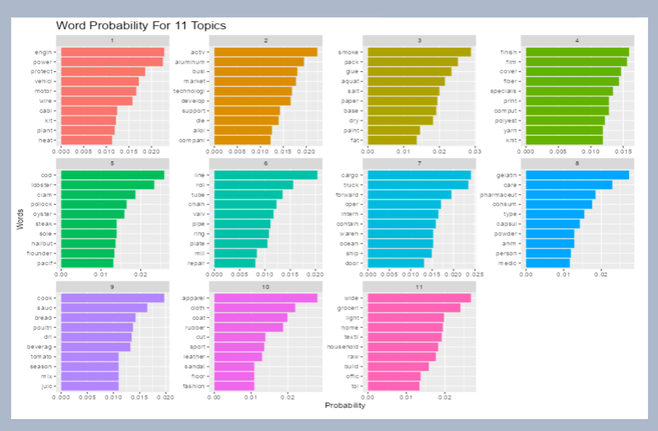


	id	country	product_services	revenue_omu
1	Jones LLC	ZH	Automobiles	310612303.447
2	Coleman, Hall and Lopez	ZH	Passenger cars, trucks, vans, and buses	162734683.9969
3	Aqua Advancements Sashimi SE Express	Oceanus	Holding firm whose subsidiaries are engaged in the businesses of refining and chemicals, process and pollution control equipment, minerals, fertilizers, polymers and fibers, commodity trading and services, forest and consumer products, and ranching	115004666.6728
4	Makumba Ltd. Liability Co	Utoporiana	Car service, car parts and accessories, automotive technology, diagnostics for repair shops, antilock braking and fuel-injection systems, auto electronics, starters, and alternators; Home (power tools for DIY enthusiasts, garden tools, household appliances, heating and warm water); and industry and trade (communication services, power tools for professional, sensors and foundry - MEMS, security systems, packaging technology)	90986412.5191
5	Taylor, Taylor and Farrell	ZH	Fully electric vehicles (EVs) and electric vehicle powertrain components	81466666.6728

Showing 1 to 5 of 8,639 entries Previous 1 2 3 4 5 ... 1,728 Next

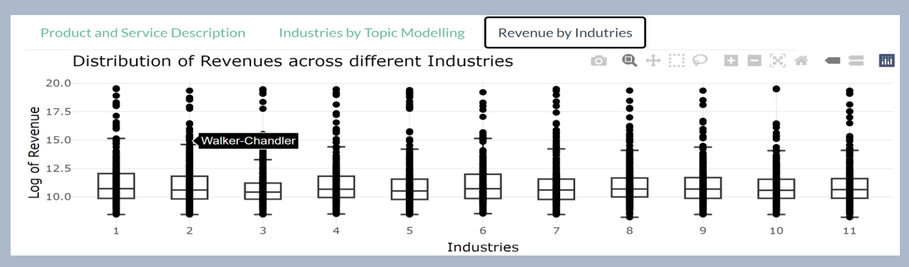
2.2 Topic Modelling to Identify Industries

Latent Dirichlet Allocation (LDA) was employed to conduct *topic modeling*, and users are offered the flexibility to choose between fitting methods such as Gibbs or Variational Expectation Maximization. This approach facilitates the determination of the optimal number of topics and enables the identification of the predominant industry associated with each topic.



2.3 Revenue Distribution by Topics

As the distribution of revenue does not follow a normal distribution, *log transformation* is applied.
A **boxplot** is generated using *ggplot* to visualize the revenue distribution across different topics, and interactivity is added using *ggplotly.*



3. Future Study

Due to the limitation of the Shiny App, we were not able to perform further analysis, to measure the similarity/differences of the different business group, such as performing the Krustal Wallis statistical test to determine if the difference in mean log revenue across different industries is statistically significant.

Besides revenue, we could also look into the business groups that operated in multiple countries to identify any anomalies.

References

[Mini-Challenge 3: (vast-challenge.github.io)]

Peer's work: Kwa Kah Boon