Vast Challenge 2023: Detect Anomalies in Fishing Business

SMU Singapore Management University
School of Computing and Information Systems

Han Shumin Wan Xinyu Wang Shengming

Advised by: Prof Kam Tin Seong

Introduction

Seafood is one of the most widely traded commodities in the global food market. More than a third of the world's population relies on fish and other seafood as a primary source of protein in their diet, and an estimated 520 million people make their livelihoods through fishing or fishing-related activities. Unfortunately, illegal, unreported, and unregulated fishing is a major contributor to overfishing worldwide. These activities pose a threat not only to fragile marine ecosystems, but also to food security in coastal communities and regional stability more broadly. The illegal fishing trade has been linked to organized crime, and human rights violations are common when fishing operations are conducted without regulatory oversight.

NGO FishEye International is a nonpartisan organization charged with understanding the social, political, and economic forces that drive the illegal fishing trade. They have spent the past several years collecting data, which they hope will help them to form a more comprehensive picture of this evolving threat. They plan to make several of their datasets available to the public, along with a series of questions surrounding illegal fishing and its broader impacts. They're asking the Visual Analytics community to help them make sense of this often-conflicting data, and to make recommendations for how to proceed.

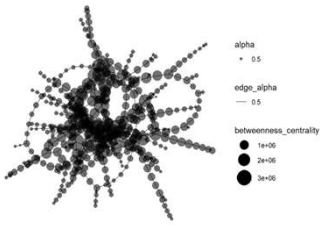
Data Preprocessing

FishEye International, a non-profit focused on countering illegal, unreported, and unregulated (IUU) fishing, has been given access to an international finance corporation's database on fishing related companies. In the past, FishEye has determined that companies with anomalous structures are far more likely to be involved in IUU (or other "fishy" business). FishEye has transformed the database into a knowledge graph. It includes information about companies, owners, workers, and financial status. FishEye is aiming to use this graph to identify anomalies that could indicate a company is involved in IUU.

1. Extracting Nodes and Edges from MC3.json File

This is an undirected multi-graph and the graph format is a json format(MC3.json) containing information of 27,622 nodes, 24,038 edges and 7,794 connected components. At the root-level, it is a dictionary with graph-level properties specified as keys (directed, mulitgraph, graph). The nodes and links keys each provide a dictionary of the nodes and links respectively. The nodes entries that must include an id key that is unique for each node. The links entries include source and target keys that refer to node id values. All other keys provided in node and link dictionaries are attributes for that node or link.

Node attributes include type of node as Defined above, country associated with the entity(this can be a full country or a two-letter country code), description of product services that the "id" node does, operating revenue of the "id" node in Oceanus Monetary Units, identifier of the node is also the name of the entry and the subset of the "type" node (not in every node attribute). Edge attributes include type of the edge as defined above, ID of the source node, ID of the target node and the subset of the "type" node (not in every edge attribute).A Brief overview of the graph network filtered by betweenness_centrality >= 100000 is as on the right.



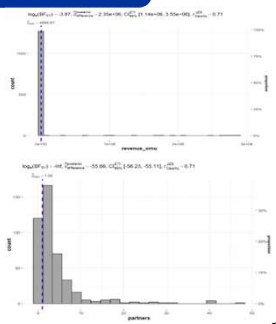
2. Study of Product service

Product service shows the relevant business the "node" does. We first do a tokenization, which is the process of breaking up a given text into units called token. After that, we count the words extracted and filter out stopwords. Then, a wordcloud is built based on the words and we filter minimum frequency of 10. In Top 10 keywords which may related to fish could be, "fish", "seafood", "frozen", "food", "fresh" and "salmon". And nodes containing these words are selected.

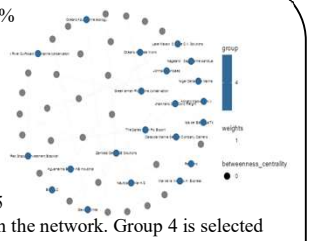
Results and Visualisation

Analysing company type

From what we discovered, the number of company type is much greater than the other 2 types, we will focus on the company type and find the distribution of revenue_omu. Most of the companies have a revenue within the first bar, but there are some companies that have far more revenue than others, we select the revenue_omu>400,000. We then calculate partner numbers (numbers of targets of a source) and assign partner = -1 if targets dont have a partner record, we only select those with a partner and group them by revenue_omu and partner numbers.

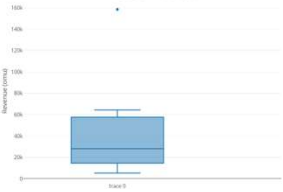


Here we define Group 1 as No. of partners > 50% and revenue <= 80%, Group 2 as No. of partners > 50% and revenue > 80%, Group 3 as No. of partners <= 50% and revenue <= 80% , Group 4 as No. of partners <= 50% and revenue > 80%, Group 5 and 6 are groups with no partners, but revenue less than or equal to 80% and revenue greater than 80%. We then visualize the nodes and edges and since Group 5 and 6 don't have partners, they will not appear in the network. Group 4 is selected since they don't have many partners in business, but they have high revenue_omu.



Communities and outliers Detection

We use a filter to extract betweenness_centrality score between 10,000 and 100,000 in order to reduce the number of nodes to be displayed in the graph. Here, the nodes with more than 3 link will be shown with label name. There is a fun fact that company owners which shown labels in the graph actually belong to different communities. Although they are running not only one companies, but they do also not have business relationships between clusters. Utilizing the groups created by community cluster detection, we will only consider the top 10 largest clusters. From the Boxplot in each cluster, we take cluster 8 as an example, we should watch out these outlier companies since their revenues are unusually higher than other companies within the same cluster.



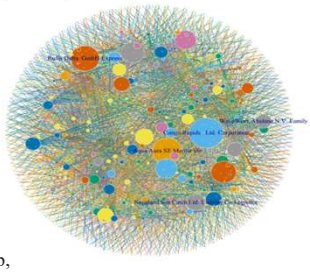
Analysing centrality and clustering

Here, we take a look at the centrality of some of the nodes with the focus on betweenness centrality as it shows the importance of nodes base on information passed through the nodes. This is important for us to identify similar business because if the node belong to a certain industry, it is highly likely that the business that communicate with it also belong to a similar industry. We have pulled out the closeness centrality as well for comparison. This centrality will be less of a focus because it measure the speed of information spreading which is less relevant for our analysis.

Show 10 entries Search:

	top_nodes_ids	betweenness_centrality	closeness_centrality
1	Congo Rapids Ltd. Corporation	1128	0.020833333333333
2	Agua Azul SE Marine life	940	0.020408163265061
3	Nagaland Sea Catch Ltd. Liability Co Logistics	923	0.010869565217391
4	WaveWorx Abalone N.V. Family	798	0.009803921568627
5	Rufiji Delta GmbH Express	780	0.025
6	Mar de la Felicidad SE	595	0.028571428571429
7	Mary Williams	594	0.010416666666667
8	Bay of Bengal's Anchor	585	0.023809523809524
9	Adair S.A. de C.V.	564	0.018867924528302
10	Spanish Shrimp A/S Marine	551	0.024390243902439

Showing 1 to 10 of 10 entries Previous 1 Next



We use the Louvain clustering algorithm to see how many communities it can detect from the graph. We have tried to use infomap, edge betweenness, walktrap clustering algorithms as well but overall results appears to be similar so let's just use Louvain for our analysis. We observe that there are 601 communities detected. This is a very large number. However, given that we have 3020 nodes in the graph and the number of attributes are limited. We will stay on with these communities. And we pick the largest 10 communities and further examine its components. We see that most of the communities with similar product_services description are grouped together. We also see that the communities includes individual personnel representing the companies as well. This might be an indication that the representatives of the companies may interact closely with the identified companies. It is also interesting to note that majority of the top 10 nodes with the highest betweenness centrality (identified earlier) belong to different communities. This might be an indication that these companies themselves are major players who may be running their own network of marine life trades

Conclusion and Future Work

In the Vast Challenge 2023 Mini Challenge 3, we use the node-link dataset provided by the FishEye International to do analysis on business groups and anomalies detection. We analyse company type, detect communities and outliers and analyse centrality and clustering to group similar business, group them and identify anomalies.

Through these initiatives, we wish to be able to provide some support to the relevant authorities in combating the frequent occurrence of illegal fishing that are happening across the oceans and do our part to identify, select and report the anomalous company to the relevant authorities such that these companies can be thoroughly checked and evaluated for such suspicious activities

We believe that our analysis and report will prove to be useful in helping to conserve the oceanic and marine life that are essential to our ecosystem.