Are Dolphins Better Than Us?

A Network Science Perspective

By: Arina Lopukhina

INTRODUCTION

Project summary:

This project analyses the dolphin community dataset as well as the Zackary Karate club dataset in order to determine whether animals and humans have similar social networks. Both datasets are unipartite and undirected.

Problem statement:

Is dolphin community more well-knit and interactive than the humans' one? Comparing and contrasting two societies.

PROJECT OUTLINE

1

EDA

- node and link count
- degree distribution
- centrality
- transitivity
- bridges

2

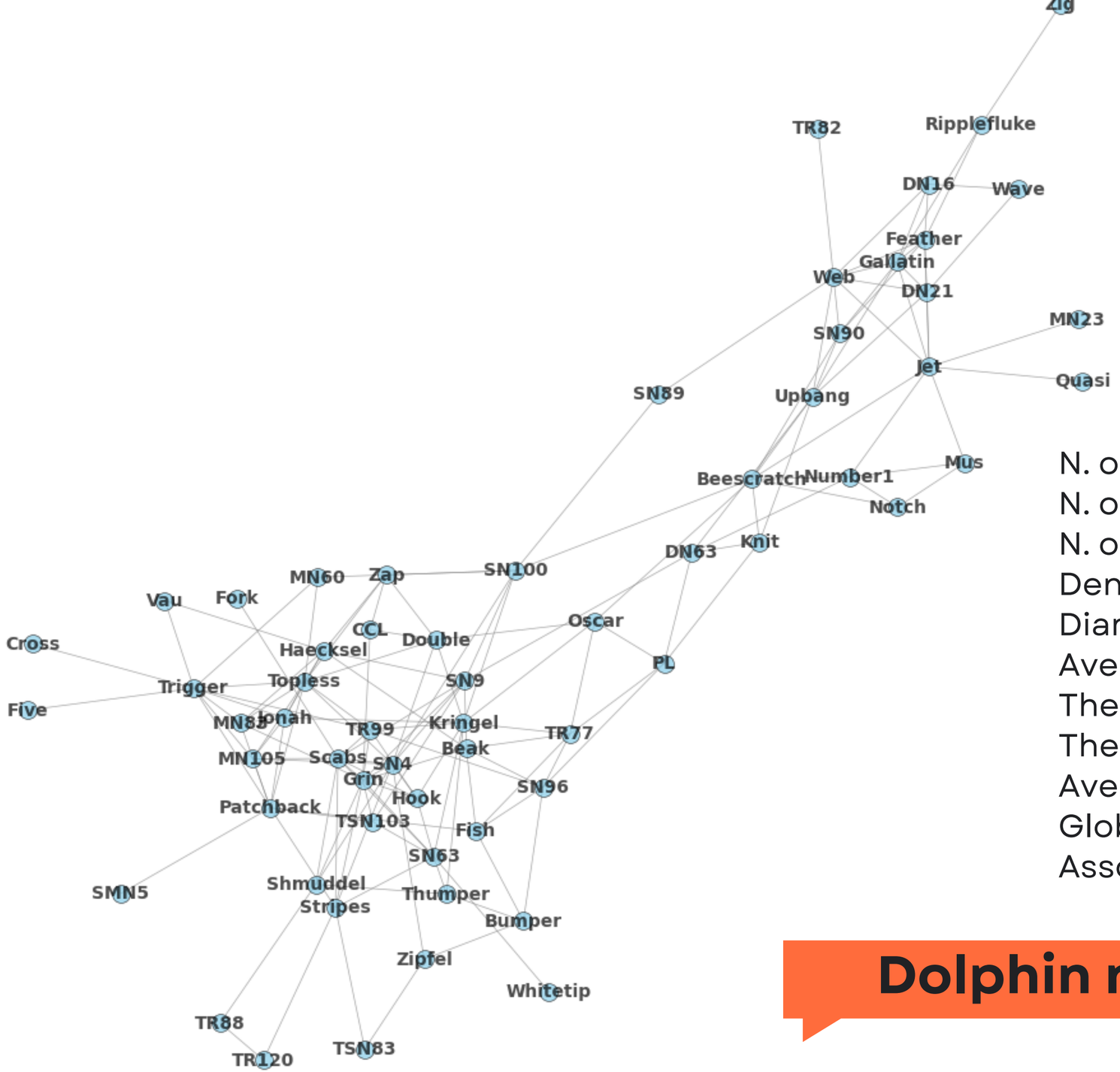
Community detection

- Greedy algorithm
- Louvain algoithm

3

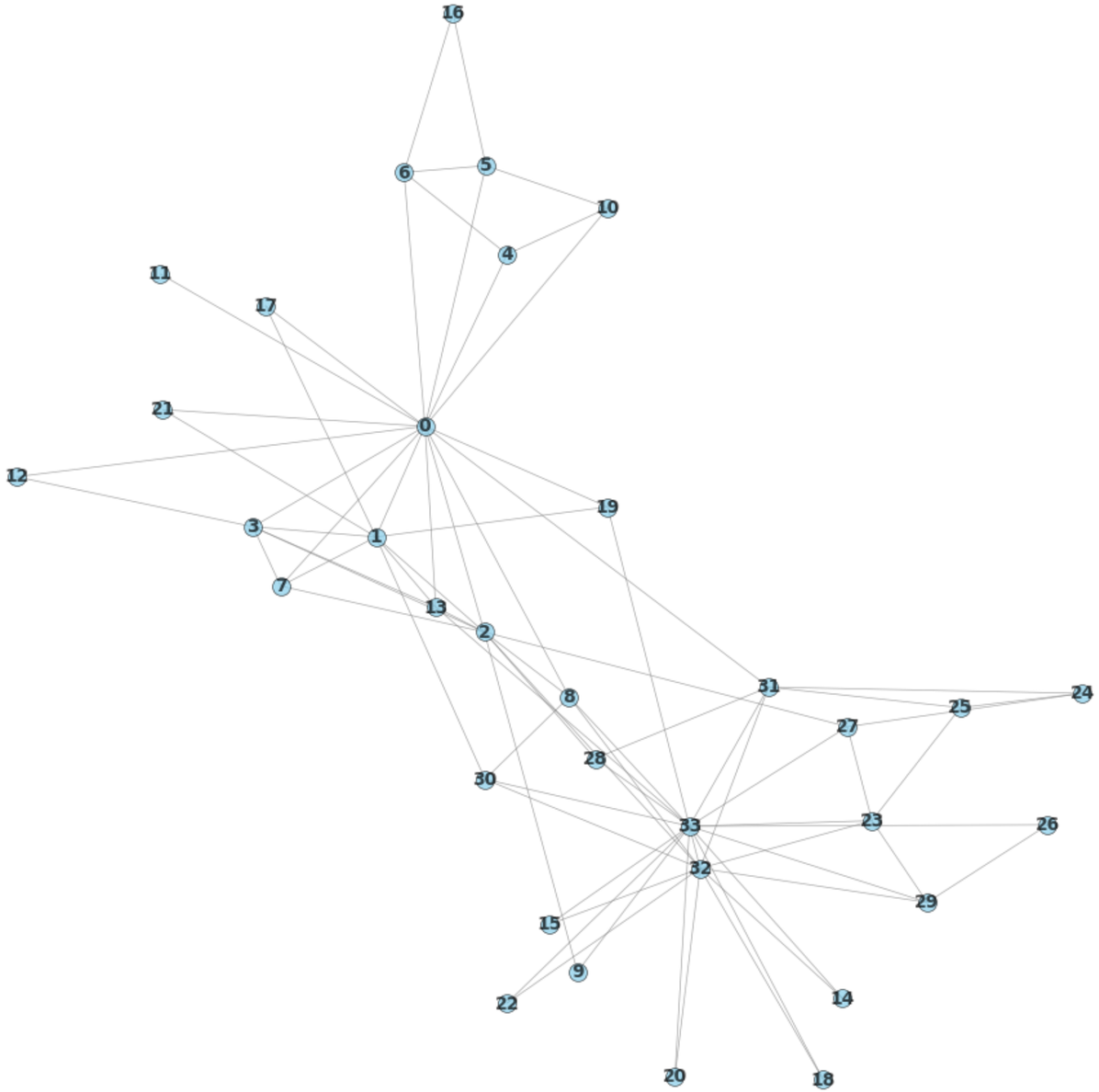
ML Link prediction

- Logistic regression
- SVM
- XGBoost



N. of Nodes: 62
N. of Edges: 159
N. of Connected Components: 1
Density of the network: **0.0841**
Diameter: 8
Average Shortest Path Length: **3.3570**
The most connected individual has degree **12**
The least connected individual has degree 1
Average Degree: **5.13**
Global Clustering Coef. (by triplets): **0.2590**
Assortativity: **- 0.04**

Dolphin network visualization



N. of Nodes: 34
N. of Edges: 78
N. of Connected Components: 1
Density of the network: **0.1390**
Diameter: 5
Average Shortest Path Length: **2.4082**
The most connected individual has degree **17**
The least connected individual has degree 1
Average Degree: **4.59**
Global Clustering Coef. (by triplets): **0.57**
Assortativity**: -0.475**

Human network visualization

Centrality information

1 Dolphin network

Top 5 highest betweenness centrality:
Node SN100: Betweenness Centrality = 0.2482
Node Beescratch: Betweenness Centrality = 0.2133
Node SN9: Betweenness Centrality = 0.1431
Node SN4: Betweenness Centrality = 0.1386
Node DN63: Betweenness Centrality = 0.1182

Top 5 highest closeness centrality:
Node SN100: Closeness Centrality = 0.4178
Node SN9: Closeness Centrality = 0.4040
Node SN4: Closeness Centrality = 0.3987
Node Kringel: Closeness Centrality = 0.3910
Node Grin: Closeness Centrality = 0.3765

Top 5 highest eigenvector centrality:
Node Grin: Eigenvector Centrality = 0.3158
Node SN4: Eigenvector Centrality = 0.3006
Node Topless: Eigenvector Centrality = 0.2850
Node Scabs: Eigenvector Centrality = 0.2811
Node TR99: Eigenvector Centrality = 0.2177

2 Human network

Top 5 highest betweenness centrality:
Node 0: Betweenness Centrality = 0.4376
Node 33: Betweenness Centrality = 0.3041
Node 32: Betweenness Centrality = 0.1452
Node 2: Betweenness Centrality = 0.1437
Node 31: Betweenness Centrality = 0.1383

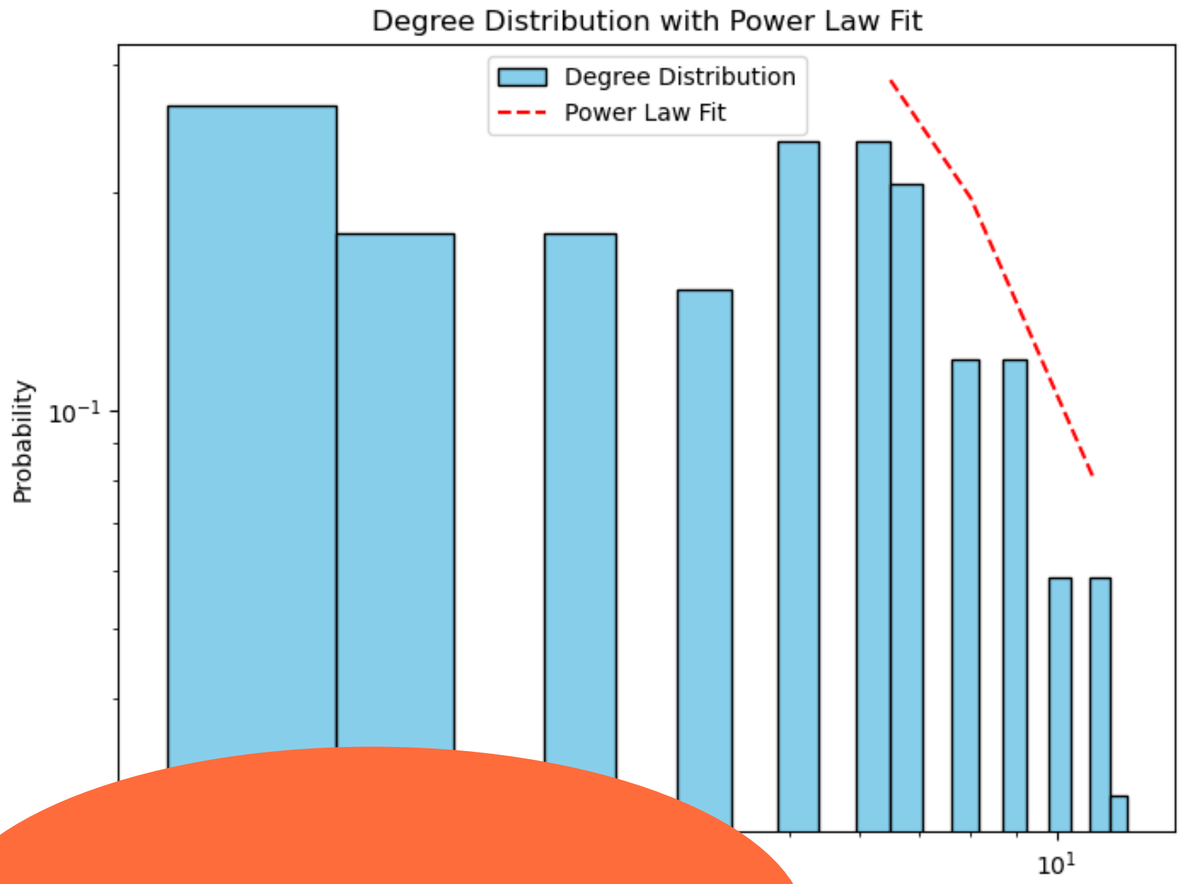
Top 5 highest closeness centrality:
Node 0: Closeness Centrality = 0.5690
Node 2: Closeness Centrality = 0.5593
Node 33: Closeness Centrality = 0.5500
Node 31: Closeness Centrality = 0.5410
Node 8: Closeness Centrality = 0.5156

Top 5 highest eigenvector centrality:
Node 33: Eigenvector Centrality = 0.3734
Node 0: Eigenvector Centrality
Node 2: Eigenvector Ce
Node 32: Eigenvecto
Node 1: Eigenvect

while human ‘leaders’ are connected, dolphins do not seem to have the same pattern

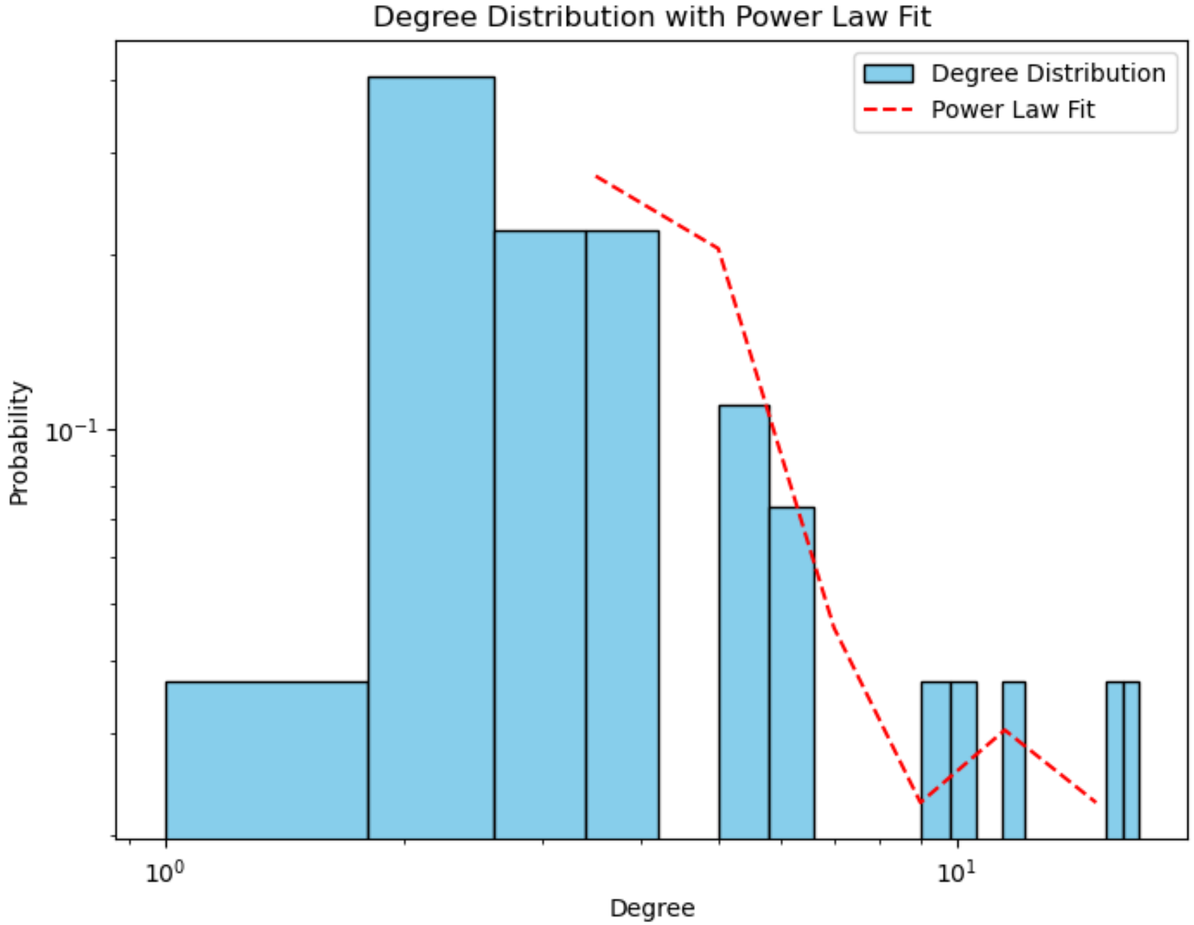
Degree distribution

1 Dolphin network



small sample size does not allow to see the distribution more clearly

2 Human network



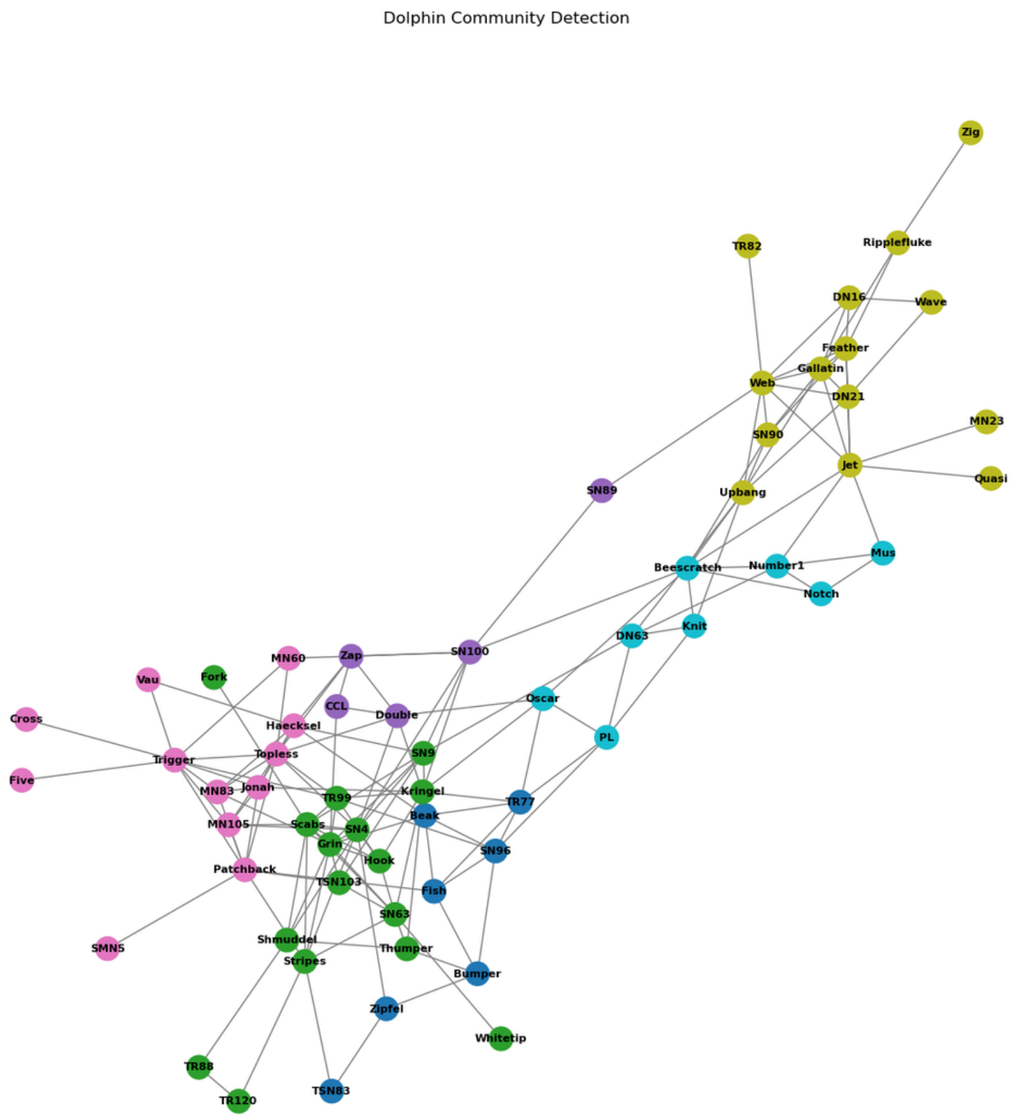
Dolphin Community detection

1 Greedy algorithm

- Network Modularity: 0.495
- Number of communities obtained by Greedy algorithm: 4

2 Louvain algorithm

- Network Modularity: 0.520
- Number of communities: 6



Human Community detection

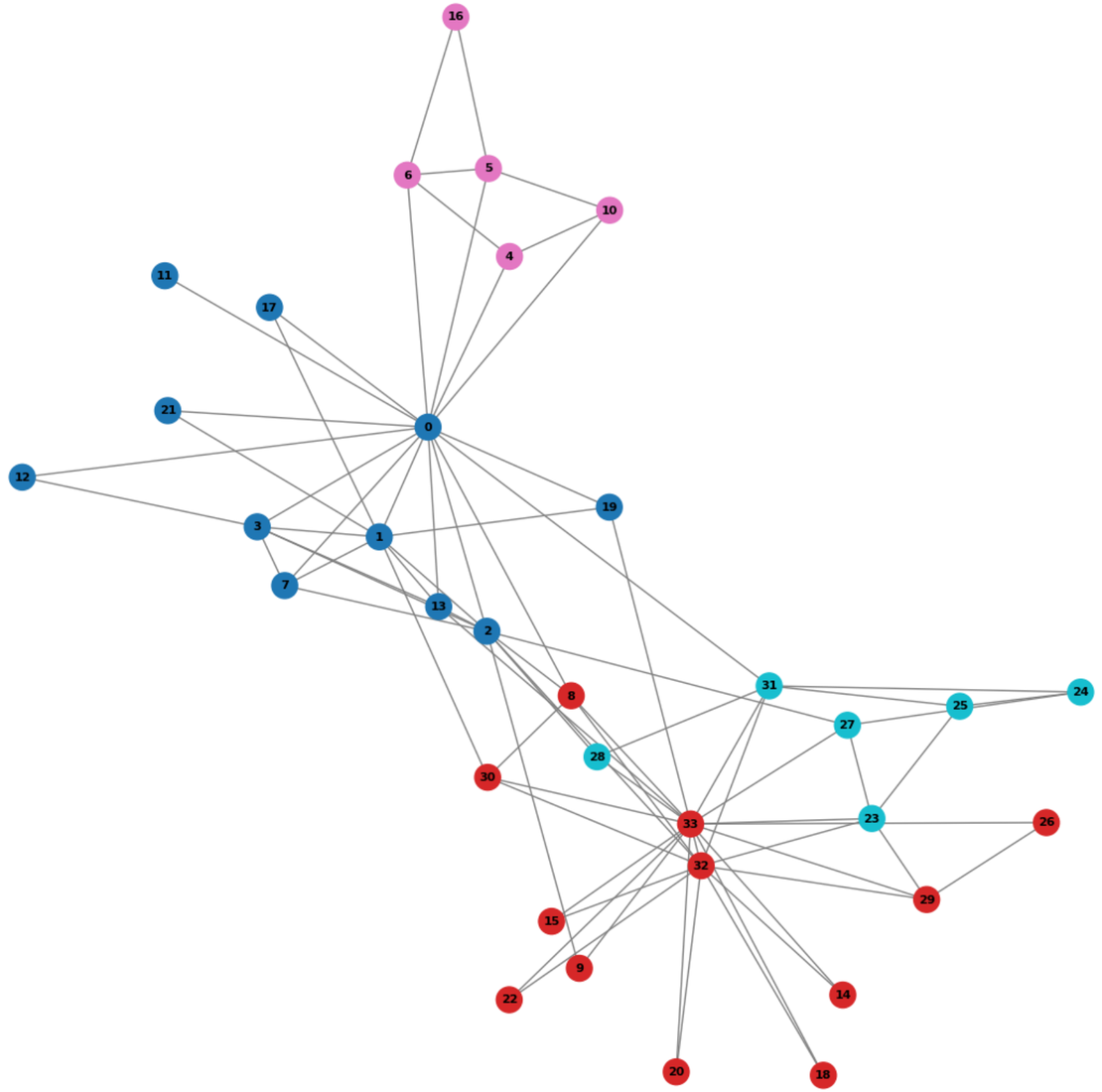
1 Greedy algorithm

- Network Modularity: 0.411
- Number of communities obtained by Greedy algorithm: 3

2 Louvain algorithm

- Network Modularity: 0.445
- Number of communities: 4

*link weights were excluded for the purpose of matching the dolphin dataset conditions.



ML Supervised Learning results

1 Aim: predicting whether two nodes would be connected

- Logistic regression
- Support-Vector Machine
- XGBoost tree classifier

+ cross-validation

2 Data preprocessing

- Assigning each node pair a label 0 or 1 depending on their relatinship
- Calculating features: Jaccard Coefficient, Resource Allocation Index, Adamic-Adar Index, Preferential Attachment Index
- Separating labels from features and splitting into test and train datasets

ML Supervised Learning results

1 Dolphin network

- Logistic regression:

Accuracy 78%, F-score 0.76, Precision 0.73, Recall 0.79

- SVM (rfb):

Accuracy: 84.38%, F1 Score: 0.84, Recall: 0.87, Precision: 0.81

- XGBoost:

Accuracy: 71.88%
F1 Score: 0.85, Recall: 0.93, Precision: 0.78

2 Human network

- Logistic regression:

Accuracy 69%, F-score 0.65, Precision 0.6, Recall 0.7

- SVM (rfb):

Accuracy: 82.38%, F1 Score: 0.81, Recall: 0.87, Precision: 0.77

- XGBoost:

Accuracy: 69.7%
F1 Score: 0.65, Recall: 0.64, Precision: 0.71

Logistic regression tended to slightly underfit, while XGBoost had overfitting problems

KEY TAKE-AWAYS

1. The human network is smaller but denser, indicating a higher proportion of possible connections are realized. Also, it has a lower average degree, shorter average shortest path length, and higher global clustering coefficient, indicating a more tightly connected and clustered structure. Both networks had a pattern with few most influential, well-connected nodes (‘leaders’), but in dolphins they did not seem to be interconnected.

2. Both networks have disassortative mixing (negative assortativity). Yet, the dolphin network was easier to split into communities, which likely mirrors the pods.

3. SVM has provided the best model outcomes for both datasets as well as consistent results, and could be potentially used for link prediction on similar datasets, while logistic regression seemed to fail to capture the data picture fully. On the other hand, XGBoost was likely too complex for the small dataset and no matter the parameter tuning it kept presenting overfitting signs.

Thank You