# A Look into the Relationships of New Zealand Dolphins through Link Analysis

Alexandras Biskis
AlexandrasBiskis@lewisu.edu

Pawel Brzek
PawelRBrzek@lewisu.edu

Victoria Griffin
VictoriaKGriffin@lewisu.edu

Joseph Griffin
JosephGriffin@lewisu.edu

DATA-51000-002 Fall 1
Data Mining and Analytics
Lewis University

# I. INTRODUCTION

Link analysis is one of the most important concepts in the field of data mining. The many algorithms related to link analysis aid in providing graph structures that represent the real world. They can analyze and find the most important links and connections (i.e. relationships) between nodes, determine the most optimal paths and network structures within the graph (e.g. maximal vs. maximum), amongst other things. The steps taken to prepare the selected dataset for analysis and the classification processes via link analysis will be discussed throughout the entirety of this paper.

The purpose of this assignment was to select a dataset capable of link analysis and perform several types of analysis on it, such as determining global qualities of the network, finding the most important nodes, and finding communities within the network. The results must be visualized to confirm results. Software and programming packages, such as Gephi, NetworkX, and SNAP, are commonly used to visualize these networks. In this analysis, Gephi, an open-source visualization package created in July 2008, will be utilized due to its user-friendly GUI and wide applicability. Gephi has various statistical functionalities on top of algorithmic and customization options which enable users to investigate datasets from many different perspectives.

To elaborate a bit more on the procedure, after preprocessing the data, the first step is to determine the global qualities of the network. This includes finding the number of vertices and edges (i.e. nodes and links), diameter, path length, average degree, density, and the clustering coefficient. All of these properties will be described at length, and then will be used to determine whether the network exhibits small-worldness, etc. After determining global qualities, the next step will be to find the most important nodes using the Betweenness Centrality, PageRank, and HITS algorithms. Further analysis of what the algorithms say about the most important nodes will be looked at carefully. Lastly, communities within the network will be found and analyzed. Each community found will be described (i.e. properties of individual communities) and labeled. Each section will be validated with visuals and thorough explanations along with those visuals.

The dataset used was posted by Mark Newman, a physics professor associated with the Department of Physics and Center for the Study of Complex Systems at the University of Michigan. He posted this dataset specifically formatted to be used with Gephi and ready for link analysis [1]. The dataset itself comes from a study conducted by a group of scientists, D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, in New Zealand studying a bottlenose dolphin community. They published their work in 2003, in the journal of Behavioral Ecology and Sociobiology, called *'The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations'* [2].

Although Betweenness Centrality, PageRank, and HITS are all algorithms used for determining link analysis, there are unique nuances for each one that would warrant further discussion. The Betweenness Centrality algorithm at its most basic is the number of shortest paths going through the nodes. It can be computed using a tree-like network structure. The betweenness centrality is computed by working its way up the tree from the bottom by counting the number of shortest paths from the one specific node to all other nodes in the network. Each path is weighted differently depending on if a node has more than one path or not. An important node will have a smaller betweenness centrality score, meaning it has close connections to the other nodes around it [5].

PageRank is a complex algorithm with lots of moving parts. Broken down to its essence, each node (or webpage as it was originally used), is ranked by importance based on weighted links from other nodes (other webpages). Which means that each node's "vote" (i.e. in-link) is proportional to the importance of the source node. The more out-links a node has, the more the rank of its vote decreases in importance. The sum of a node's in-links will determine its importance. A node's importance increases significantly if it is pointed to by other important nodes. To find the most important nodes in a community, this requires some linear algebra in which computation continues until a convergence eigenvector is revealed. The nodes with the largest values in the eigenvector will have the highest importance [3].

The HITS algorithm is similar to PageRank and was proposed around the same time, about 25 years ago. The HITS algorithm is ranked by two unique scores; the hub score (i.e. the quality as an expert, meaning the sum of votes of authorities it points to), and the authority score (i.e. quality of content, meaning the total sum of votes coming from other experts). Similar to PageRank, the algorithm then uses these two scores, employs linear algebra, and iterates through until it converges. Ultimately what makes HITS different from PageRank, is that PageRank's algorithm values the most important nodes by focusing on in-links, and HITS algorithm values the most important nodes by focusing on out-links. Careful attention was paid to the HITS data within this paper, as it can provide more information for us than PageRank [4].

The sections of this report cover different stages of the analysis. Within Section II, the dataset is described more in detail with information ranging from the types of variables, distributions of variables, and other significant information. Section III explains the methodology of the analysis in detail and the tools utilized to carry it out. Section IV dives into the results from the tuned hyperparameters of each SML model and discusses the findings. Section V wraps up the analysis with final thoughts and future recommendations/implications.

## II. DATA DESCRIPTION

The dolphin dataset has a total of 62 nodes (i.e. vertices) and 159 links (i.e. edges). Each node represents an individual dolphin, whereas each link represents associations/relationships to other dolphins. The dataset was collected and published in 2003 [2]. Table 1, shown below, displays the most important global qualities, their values, and a quick explanation of what the values mean. The global qualities included in Table 1 are: number of nodes, number of links, diameter, average path length, average degree, density, and clustering coefficient. A more in-depth look at these will be discussed later in sections III, methodology, and section IV, results and discussion. Fig. 1, below, gives an initial look at the dataset first uploaded to Gephi with each node (i.e. dolphin) and link shown.

*Table 1. Global Qualities of the Dolphin Dataset*

| **Global Quality** | **Value** | **Quick Explanation** |
|---|---|---|
| Number of Nodes | 62 | This represents the number of dolphins studied. These 62 dolphins live in one community |
| Number of Links | 159 | This represents the number of links between the community of 62 dolphins |
| Diameter | 8 | Shortest distance between the two most distant nodes in the network |
| Average Path Length | 3.3569 | Average graph-distance between all pairs of nodes |
| Average Degree | 5.129 | Total number of edges connected to a particular vertex |
| Density | 0.084 | Measure of how close a network is to complete. Complete = 1 |
| Clustering Coefficient | Clustering Coefficient: 0.303<br>(Total triangles: 95) | Indication of a “small-world” effect. The average clustering coefficient indicates how nodes are embedded in their neighborhood. |

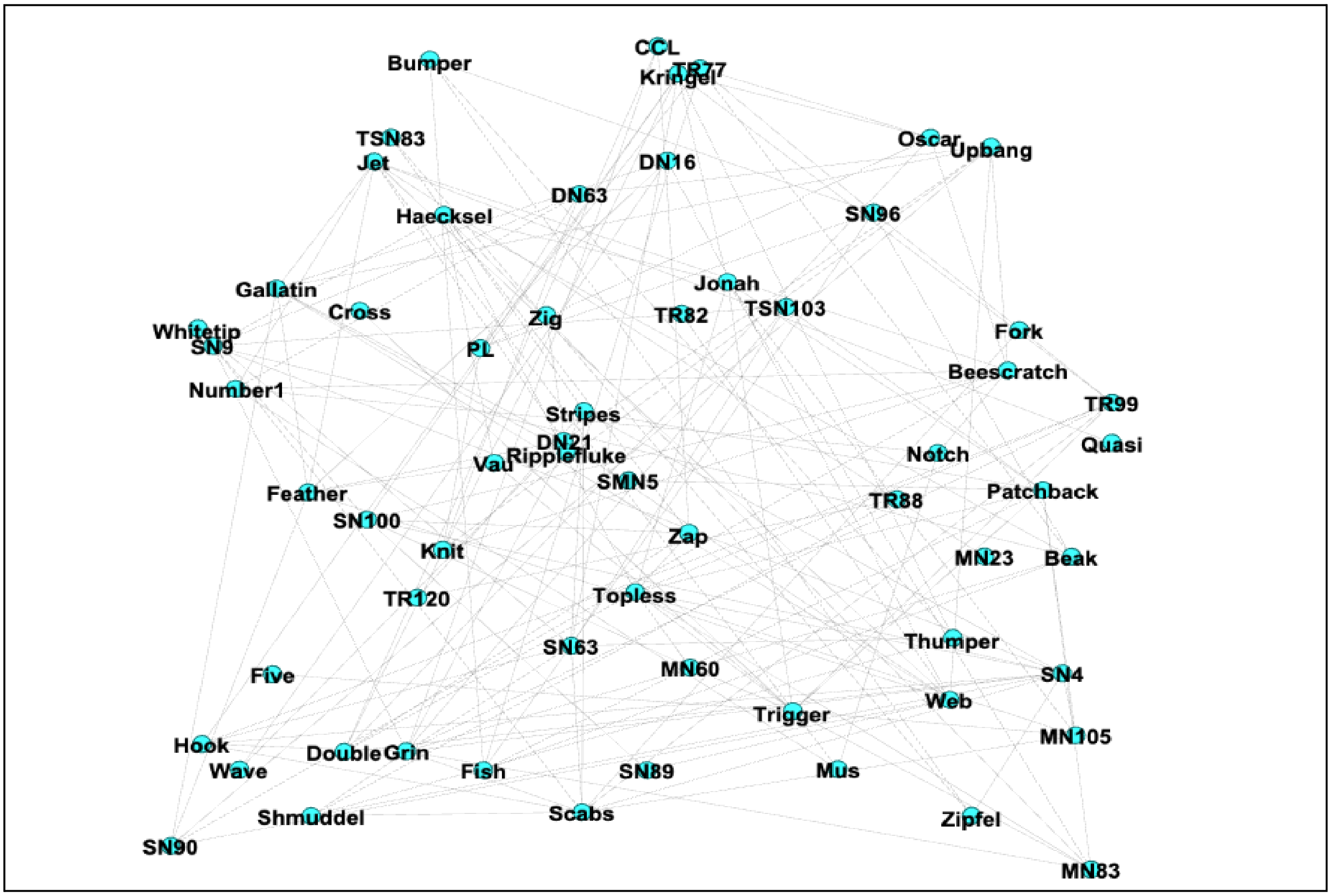

*Fig. 1. Initial network with 62 dolphins shown*

## III. METHODOLOGY

The methodology applied on the dataset for this analysis was straightforward and summarized in Fig 3. The dataset obtained was already formatted for algorithmic application through Gephi, so preprocessing was not needed. As stated from the introduction, Gephi has various statistical functionalities on top of algorithmic and customization options which enable users to investigate datasets from many different perspectives. Gephi was used for the entirety of this link analysis process.

Initially, the following global qualities were looked at to get an idea of the characteristics of our dataset: number of nodes, number of links, diameter, average path length, average degree, density, and clustering coefficient. Global qualities of a dataset can give a deeper understanding of the relationships in a dataset. It is important to first know how many nodes (i.e. vertices) and links (i.e. edges) are a part of the set. This helps to determine if the scale of the network is reasonable for introductory link analysis. Our dataset has 62 nodes and 159 links (as seen in Table 1), which was agreed to be a reasonable size for this analysis.

Diameter was the next global quality found. The diameter of the graph is found in two steps. First, the shortest path between two pairs of nodes for each possible pair in the network is found. Then from that data, the greatest length found between any two pairs will be recorded as the diameter. For this dataset, the diameter was 8, meaning the longest distance (of all the shortest paths found) was no more than 8 links apart.

The next global quality was finding average path length. To find the average path length, all of the shortest distances between all possible pairs of nodes are summed and divided by the total number of pairs as seen in Equation (1). Within the equation, n represents the possible pairs, and $\text{nu}_1$ and $\text{nu}_2$ and the node pairs. The average path length of this dataset was 3.3569.

$$l_G = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \neq j} d(v_i, v_j) \tag{1}$$

*Equation 1. Mathematical Equation for Average Path Length*

After average path length, the average degree was found. The average degree is calculated by taking the total number of links and dividing by the total number of nodes. Equation (2), below, reiterates average degree.

$$\text{Average Degree} = \text{Total \# Links/Total \# Nodes} \tag{2}$$

*Equation 2. Mathematical Equation for Average Degree*

Next, network density was found. Network density is specifically a measure of the network's "health", meaning the portion of connections in a network that are actual connections (i.e. number of links). Potential meaning the ones that could exist, versus the connections that actually exist. Equation (3) and equation (4), shown below, shows how to first calculate the potential connections, and then how to calculate network density. Our network density was 0.084.

$$\text{Potential Connections} = \frac{\text{Nodes} * (\text{Nodes} - 1)}{2} \tag{3}$$

*Equation 3. Mathematical Equation for Potential Connections*

$$\text{Network Connections} = \frac{\text{Actual \# of Connections}}{\text{Potential \# of Connections}} \tag{4}$$

*Equation 4. Mathematical Equation for Network Density*

The last global quality that was observed was the clustering coefficient. The clustering coefficient is essentially the degree to which the nodes cluster together in their communities. If the network/community is fully connected, the clustering coefficient will be close to 1, and a clustering coefficient of 0 would mean hardly any connections in the community. How we calculate that is the number of closed triplets (i.e. closed triangles), over the total number of triplets, both open and closed. Equation (5) shows how to calculate the global clustering coefficient. Our clustering coefficient was 0.303 (95 total triangles).

$$CC = \frac{3 \times \text{number of triangles}}{\text{number of triplets}} = \frac{\text{number of closed triplets}}{\text{number of triplets}} \tag{5}$$

*Equation 5. Mathematical Equation for Clustering Coefficient*

Arguably the most important part of a network analysis such as this is to identify important vertices or nodes. The 'importance' of a node can give valuable information into the structure of the network itself. The common approach to this identification process is to assess the centrality of various nodes. Centrality essentially means ranking nodes based on their structural importance within the network that they reside in. Centrality can be assessed through various methodologies, however, this analysis only focuses on three such processes and they are as follows: betweenness centrality, PageRank, and Hyperlink-Induced Topic Search (HITS).

Betweenness Centrality was the first measure of centrality used in this analysis. Betweenness Centrality is based on the extent, or the number of times, to which a particular node is part of a certain shortest path between other nodes. At its essence, betweenness centrality indicates which nodes serve as a so-called 'bridge' to other nodes in the network. In other words, it is commonly used to understand which nodes influence the flow of a network the most. Equation (6) demonstrates how to calculate the betweenness centrality of a node using $\sigma_{st}$ as the total number of shortests from paths from node s to node t, and $\sigma_{st}(v)$ as the number of the shortest paths that pass through node v.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{6}$$

*Equation 6. Mathematical Equation for Betweenness Centrality for a Node*

PageRank was the second measure of centrality used in this analysis. PageRank, originally developed by Google to rank websites, is readily applicable in social networks such as this one. PageRank looks at the important nodes that have many 'in-links' from other high ranked nodes. The output of the PageRank algorithm represents the likelihood of ending up at a certain node when you start at any given node. Equation (7) demonstrates how to calculate the PageRank of node u using each PageRank value for each node v divided by the number of links from node v.

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \tag{7}$$

*Equation 7. Mathematical Equation for PageRank for a Node*

HITS was the third measure of centrality used in this analysis. HITS is also known as the combination of hubs and authorities. HITS uses this combination to explain the recursive relationships that might be seen between certain nodes. The idea behind using authorities and hubs is to assess whether a node is of high-quality, or relevance. Authority is an 'in-link' value where a node is categorized as high-quality if many other high-quality nodes also link to it. Hub is an 'out-link' value where a node is categorized as high-quality if the node in question links to many other high-quality nodes. This representation of 'in-link' and 'out-link' can be seen in Fig 2.

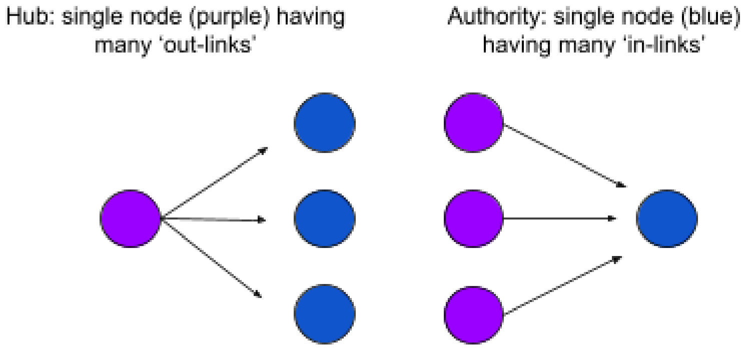
Hub: single node (purple) having many 'out-links'
Authority: single node (blue) having many 'in-links'

*Fig. 2. Representation of Hubs and Authorities*

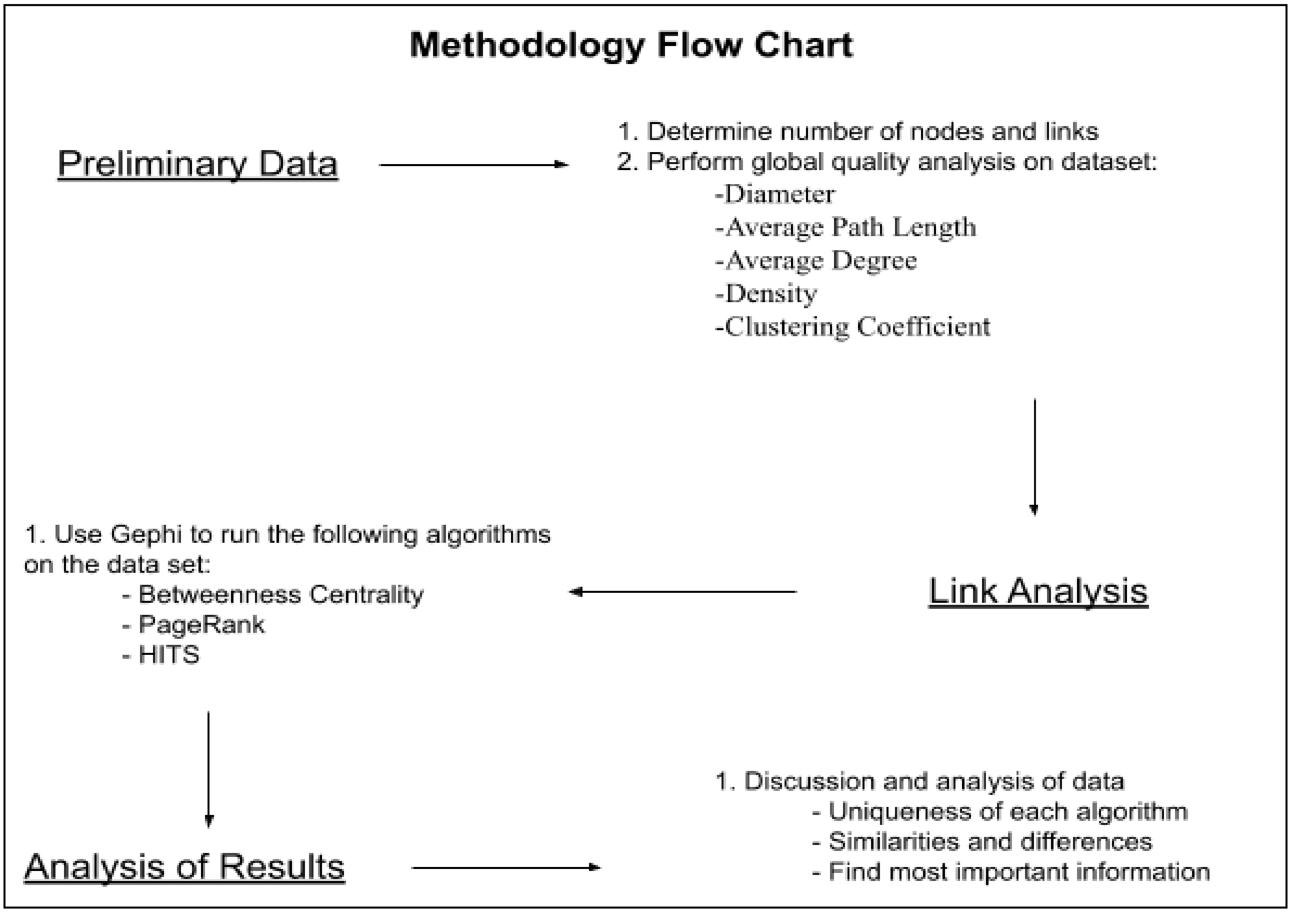
Methodology Flow Chart
Preliminary Data
1. Determine number of nodes and links
2. Perform global quality analysis on dataset:
-Diameter
-Average Path Length
-Average Degree
-Density
-Clustering Coefficient
Link Analysis
1. Use Gephi to run the following algorithms on the data set:
- Betweenness Centrality
- PageRank
- HITS
Analysis of Results
1. Discussion and analysis of data
- Uniqueness of each algorithm
- Similarities and differences
- Find most important information

*Fig. 3. Methodology Schema Layout*

## IV. RESULTS AND DISCUSSION

There are certain global qualities in particular that can be used to dig deeper into the health of the dolphin community. Average path length and the clustering coefficient aid in determining what is called the small-worldness property. The dolphin's network average path length was 3.3569 (under the ln(|node|) value of 4.1271), which is a good sign for the connectedness of the network. The dolphins community is such that all of the dolphins are closely involved with the other dolphins. Although this is true, the clustering coefficient was low, just 0.303 (maximum being 1). So although these dolphins live in a small tight-knit community, they all don't interact or have relationships with each other. This makes us think that many of these dolphins might be babies or very young, where they will only interact with very few other dolphins until they get older.

As mentioned in the previous section of the document, Betweenness Centrality allows for the detection of how influential a node can be in the flow of the information or indicate if a node or nodes hold authority over disparate clusters (sub-groups) in a network.

The results of the calculation on the project dataset for the Betweenness Centrality algorithm indicated that the dataset could be broken into seven centrality clusters that are made roughly of around 10 nodes per the cluster with the following Betweenness Centrality value (BCV) ranges for each cluster:

- Cluster 1 – BCV values of 0.0
- Cluster 2 – BCV values between 0.000137 and 0.009073
- Cluster 3 – BCV values between 0.012038 and 0.029373
- Cluster 4 – BCV values between 0.032695 and 0.057166
- Cluster 5 – BCV values between 0.061972 and 0.099122
- Cluster 6 – BCV values between 0.1143 and 0.14315
- Cluster 7 – BCV value higher than 0.2000

With the above information the results indicate two (2) nodes that standout from the reset of the nodes in the dataset with values above the level of 0.20 (Cluster 7), followed by five nodes with the values above the level of 0.10 (Cluster 6), as seen in Table 2.

*Table 2. Betweenness Centrality Results (Partial) for Dataset*

| **Node** | **Value** | **Quick Explanation** |
|---|---|---|
| SN100 | 0.248237 | Node with the highest Betweenness Centrality located in cluster 7 |
| Beescratch | 0.213324 | Second highest Betweenness Centrality value located in cluster 7 |
| SN9 | 0.14315 | Highest value node from the Cluster 6 |
| SN4 | 0.13857 | Second highest value node from the Cluster 6 |
| DN63 | 0.118239 | Third highest value node from the Cluster 6 |
| Jet | 0.1143 | Fourth highest value node from the Cluster 6 |
| Kringle | 0.102646 | Fifth highest value node from the Cluster 6 |

Based on the above results along with the visualization of the network, seen in Fig 4, and taking into the consideration the type of the network information that the dataset represents along with the evidence on the dolphins' societies [6]: highly social, compromised of distinct subgroups; the data appears to support the notion of a highly social network of dolphins made up of several clusters (subgroups): females with calves, juvenile subgroups and bachelor groups; that are made up of with distinct individuals that play key roles in network hierarchy and with two dolphins leading this specific group of dolphins.

As a possible explanation of the data, Beescratch and NP100 might be the alpha dolphins, the oldest, the most social, or the main breeders. Dolphins with a score of 0 are the least social in the group. Maybe these are babies, young dolphins, or social rejects. Dolphins with the highest scores seem to have the most rich and widespread relationships with the other dolphins in the network. They are the heart of their network.

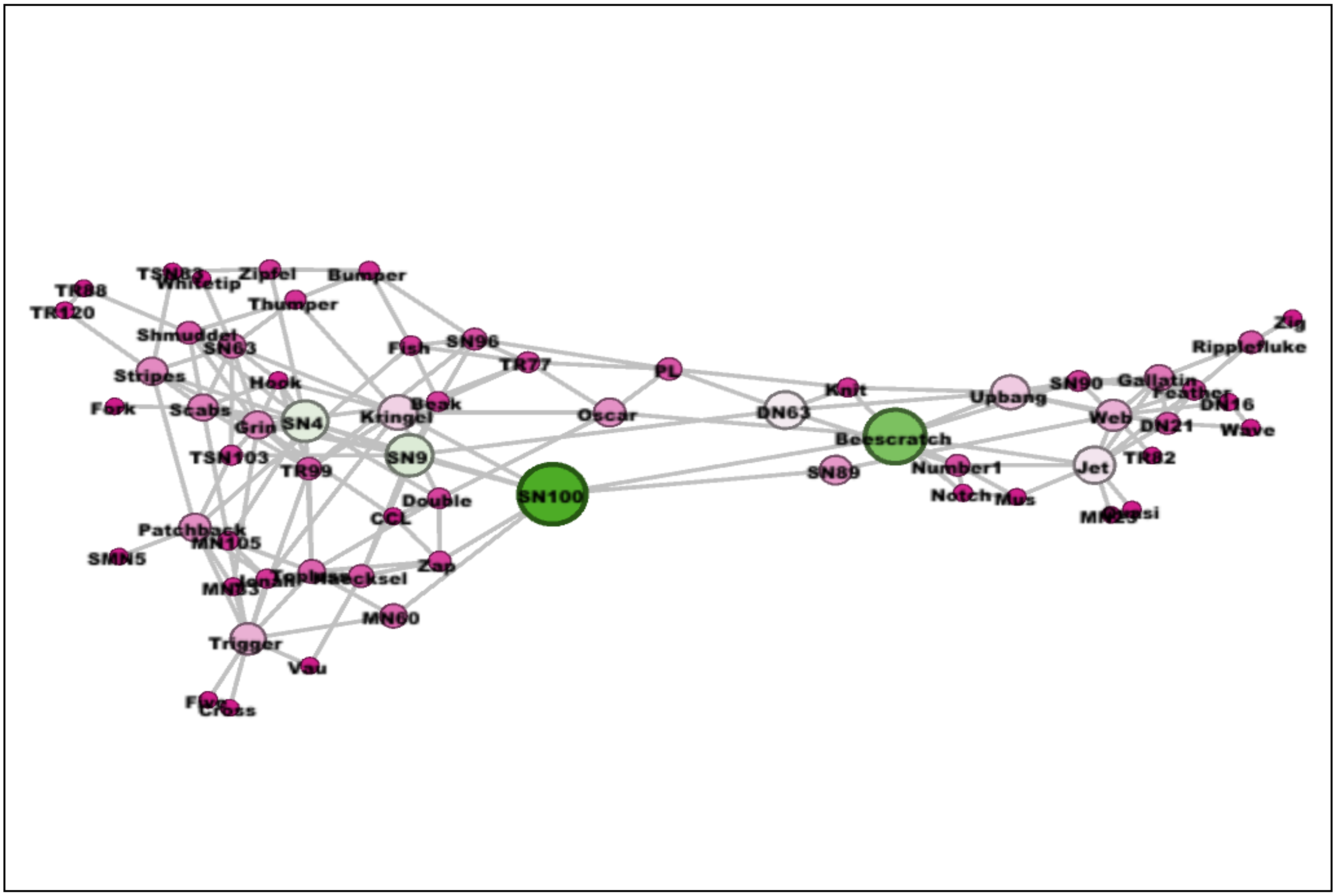

*Fig. 4. Dataset Visualization of Betweenness Centrality*

The PageRank algorithm ranks what nodes have many ‘in-links’ from other highly ranked nodes. Depending on how many ‘in-links’ a node has, and the weight of those ‘in-links’, the PageRank algorithm shows the likelihood of ending up on that given node. Table 3 shows the top six dolphins and their PageRank scores. For reference, the average PageRank score for all nodes was 0.016129. Below, Fig. 5 shows the six dolphins that had the highest PageRank scores represented by the larger dark-blue nodes. These six dolphins can be thought of as the members of the community who have the most direct ties to other dolphins within the population.

The richness of those links give us hints as to what relationships they might have with the others. With how many in-links they have, and the weight of them, these dolphins might possibly be the breeding dolphins. Without these important links, the dolphin community would not survive. The next level of weighted in-links, represented by the medium and light-blue nodes, could very well represent the breeding females.

*Table 3. Dataset Node PageRank*

| **Node** | **PageRank Value** |
|---|---|
| Grin | 0.032 |
| Jet | 0.032 |
| Trigger | 0.031 |
| Topless | 0.030 |
| Web | 0.030 |
| SN4 | 0.030 |

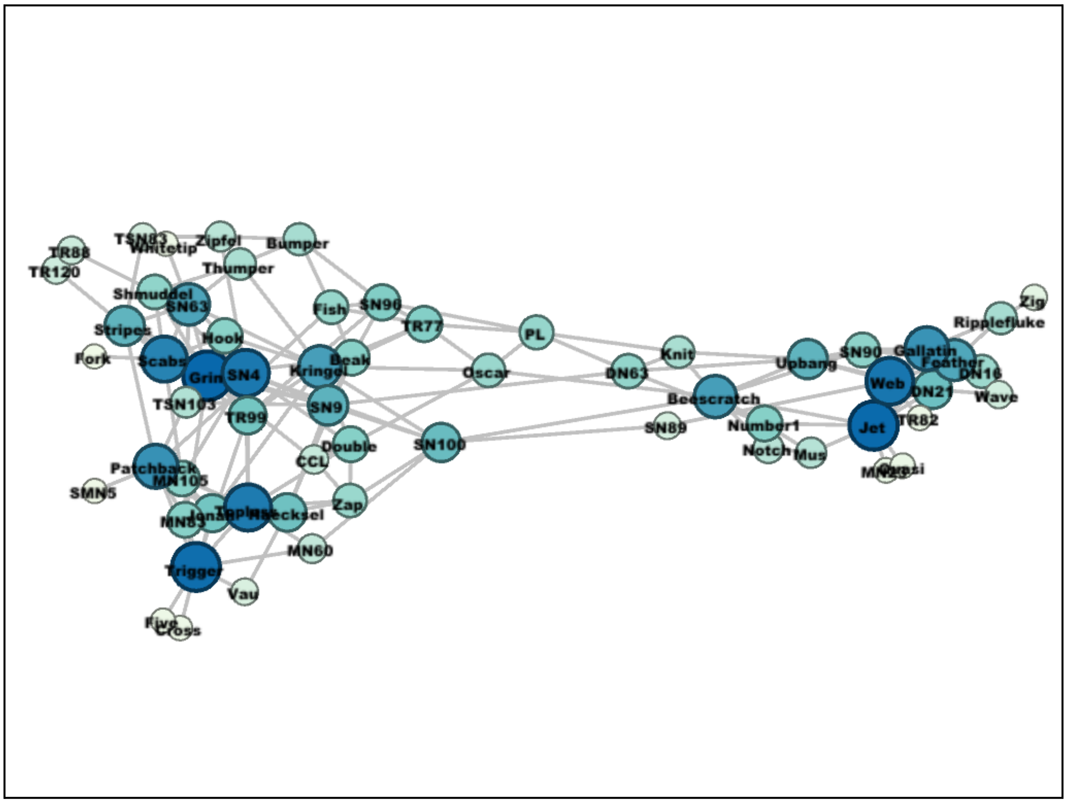

*Fig. 5. Dataset Visualization of PageRank*

As described previously, the main values that can be observed from the HITS algorithm are the hub and the authority scores. Looking at these two values together allows for a better understanding of the roles for each of these nodes. Opposite to PageRank, the HITS algorithm focuses on the importance of the 'out-links' (i.e. its hubs) as well as its 'in-links' (i.e. its authority). Shown in Table 4, the following six dolphins had the highest HITS scores.

*Table 4. Dataset Node HITS*

| **Node** | **HITS (Authority/Hub)  Values** |
|---|---|
| Grin | 0.315783 |
| SN4 | 0.300562 |
| Topless | 0.285005 |
| Scabs | 0.281099 |
| TR99 | 0.217692 |
| Patchback | 0.211763 |

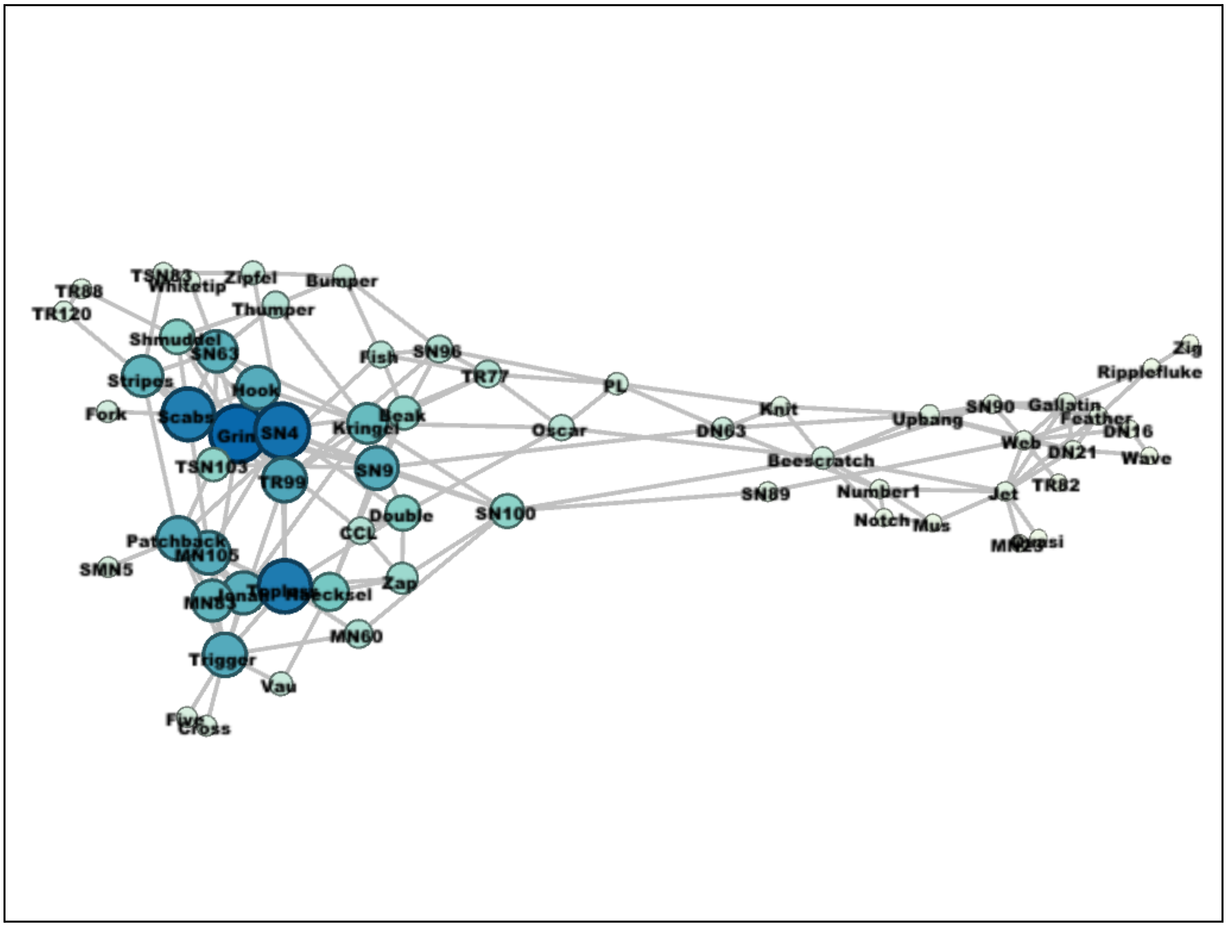

*Fig. 6. Dataset Visualization of HITS*

Furthermore, the scores on average are much greater in the left community than in the right community, as denoted by the darker blue nodes. From the three link analysis algorithms used, there are possibly two main communities. Betweenness Centrality gave us two main dolphins that may be the heart of these two communities (e.g. oldest or the alphas). Based on the other two algorithms, SN100 is the alpha of the older, more established community (left-side) and Beescratch is the alpha of a newer, 'extended family' (right-side) community. From there, six dolphins were found from both the older and newer communities highlighted as the top scores in the PageRank algorithm. These six dolphins from both communities (having the most ‘in-links’) are thought to be the top breeders. Their importance is to grow the community and strengthen its size. HITS helps us put it all together, by hinting possibly at the age of the communities, via focusing on the ‘out-links’. The highest scores in HITS highlight the older community (left-side) much more than the younger community (right-side). ‘Small-worldness’ was not observed in this network, as seen through the relatively small clustering coefficient. This would perhaps suggest that a majority of the dolphins are of a younger age and do yet have well established connections throughout the community that may be observed with older dolphins.

## V. CONCLUSION

As stated previously the goal of this assignment was to select a dataset capable of link analysis and performing several types of analysis on the said dataset, such as determining global qualities of the network, finding the most important nodes, and finding communities within the network. Resulting information was then visualized to confirm results of the calculation by different algorithms. To help achieve the purpose of the software and programming packages, such as Gephi, NetworkX, and SNAP, commonly used to visualize these networks. In this analysis, Gephi, an open-source visualization package created in July 2008, was utilized due to its user-friendly GUI and wide applicability. Gephi has various statistical functionalities on top of algorithmic and customization options which enable users to investigate datasets from many different perspectives.

It was not the methodology that took up a majority of time when achieving the purpose of this paper, but rather the analysis of our global quality, betweenness centrality, PageRank, and HITS data. The most important aspects of each, the differences and similarities, and ultimately which algorithm did the best job required deep thought and discussion from the entire group. The group is confident in the analysis of the complexity and health of the dolphin dataset used for this paper.

This community is complex and thriving as can be seen with the results from the different algorithms. The most important dolphins sometimes changed depending on how the algorithm weighted the links and nodes, denoting that there are many different ways to look at a network. The small worldness gave the first hint into what might be seen when applying the three link analysis algorithms; that the community has a short path length, meaning the dolphins all are socially connected in a meaningful way. But the clustering coefficient was low, meaning not all possible connections that could be, exist. This community might possibly be young (many being babies or children), which could be a valid explanation for the short path length, but low clustering coefficient.

In the future, analyses should consider looking at inter-cohort dolphin populations from neighboring areas as well, not just off the coast of New Zealand, to gain an even more extensive understanding of how different dolphin populations may or may not interact with each other. Dolphins are intelligent mammals with highly evolved social skills and migratory patterns, so this idea is not overtly farfetched.

# REFERENCES

[1] M. Newman, “Dolphin social network: an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand”, 2003. [Online]. Available: http://www-personal.umich.edu/~mejn/netdata/

[2] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, “The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations,” *Behavioral Ecology and Sociobiology* 54, 396-405, 2003 [Online] Available: https://link.springer.com/article/10.1007/s00265-003-0651-y

[3] L. Page, S. Brin, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” *Computer Science Department. Stanford University Stanford. CA,* 1998 [Online] https://snap.stanford.edu/class/cs224w-readings/Brin98Anatomy.pdf [Accessed: October 18, 2022]

[4] J. Kleinberg, “Authoritative Sources in a Hyperlinked Environment,” *ACM-SIAM Symposium on Discrete Algorithms,* 1998 [Online] https://www.cs.cornell.edu/home/kleinber/auth.pdf [Accessed: October 18, 2022]

[5] L. Freeman, “A Set of Measures of Centrality Based on Betweenness,” *Sociometry, Vol 40, Number 1, 35-41,* 1977 [Online]. Available: https://www.jstor.org/stable/3033543

[6] Dolphin Research Center, Dolphin Culture [Online] https://dolphins.org/culture [Accessed: October 19, 2022]