Primate Grooming: Understanding Social Social Structure Through Link Analysis

Joe Cruz
josephbcruz@lewisu.edu
DATA-51000-001, Summer 1 (2021)
Data Mining and Analytics
Lewis University

I. Introduction

Ethology, or the study of animal behavior, is a very diverse field that affects a wide variety of industries and aspects of society. Namely, ethology has special implications in topics such as scientific research (animal studies), animal husbandry (livestock and farm animal breeding/caregiving), and environmental or biological conservation. In the application of biological conservation, understanding animal behavior could lead to an understanding of infectious disease spread in a given population. More specifically, by understanding animal behavior it is possible to understand how the infectious diseases spread, whether it through explicit physical contact or potential exposure to excreta. Furthermore, with respect to animal studies, it is incredibly important to understand the animals' behavior to determine a baseline for the study. For instance, if one were to utilize a primate in a study for the development of a depressant, assuming basic social behaviors of the primate were understood, it would be much easier to identify abnormalities in primate's behavior. This basic understanding may ultimately lead to less speculation when determining the causation of the behavior change.

Social behaviors of animals are especially important to understand because the behaviors between individuals of a species can demonstrate community values, instinctual responses, and, potentially, deviations from the norm of the rest of the species. Primates are incredibly social creatures that have a myriad of different social behaviors. One of these many behaviors is grooming, where an individual cleans another individual by removing dirt, insects, and tangled fur to help maintain the individual's skin and hair conditions [1]. However, it is possible that there are other societal implications to grooming other than hygiene. To understand these implications, it is important to observe them and determine potential relationships between the actions and their frequencies with an individual. To do this, one can utilize a link analysis to extract information from social interaction observations. By extracting information from the link analysis, it could be possible to determine potential rationales for the social interactions, determine community structure with the frequency of interactions, or determine individual relationships within the community to other individuals through the frequency of interactions.

In this paper, a link analysis is performed to determine potential relationships found in *Macaca fuscata* grooming. This analysis will be performed utilizing a dataset containing grooming observations and frequencies between individual *Macaca fuscata* [2,3]. Throughout this report, a description of the dataset utilized will be explored as well as a summary of the methodology utilized for analysis. Moreover, a results and discussion section will be included along with a conclusion section from the data shortly after. In Section II of this report, the dataset will be summarized in detail. Section III will contain the summary of the methodology utilized for the link analysis. In Section IV, the results of the analysis as well as a subsequent discussion will ensue to clarify and draw conclusions from the results. Finally in Section V, the conclusions of the findings will be presented.

II. Data Description

The dataset that will be used for this link analysis is a dataset originally from a study regarding the spread of infectious disease in primate social networks [2, 3]. This data was obtained on a free-ranging *Macaca fuscata* in Yakushima, Japan. Further, the dataset specifically recorded instances of the observed grooming interactions between individual *Macaca fuscata*. There are a total of 1340 observations recorded (instances) in the dataset with a total of 4 features, which are presented in Table 1. The first feature of the dataset is "ID_of_groomer", which is the numeric identifier of the *Macaca fuscata* observed performing the grooming. The next feature of the dataset is "ID_of_groomed", which is the numeric identifier of the *Macaca fuscata* observed receiving the grooming. The next attribute is "Edge_weight", which is the frequency of the association between the two Macaca fuscata observed. In other words, this is the frequency in which this is observed between the two individuals for that day. Finally, the last attribute of the dataset is "Edge_Timestamps", which represents the day that the observation took place. Given that the data was collected over a 3-month time frame, the timestamps are in terms of the days. Of the dataset, all four attributes were utilized during the link analysis. The frequency distributions of the "ID_of_groomer", "ID_of_groomed", and "Edge_Timestamps" features can be found in Figures 1-3. For the "ID_of_groomer" and the "ID_of_groomed" attributes, there are a total of 25 primates that were observed performing

TABLE I. ATTRIBUTES OF THE PRIMATE ASSOCIATION DATASET

Attribute	Type	Example Value	Description
ID_of_groomer	Numeric (string)	19	Identification number of *Macaca fuscata* either grooming or being groomed.
ID_of_groomed	Numeric (string)	5	Identification number of *Macaca fuscata* either grooming or being groomed.
Edge_weight	Numeric (decimal)	0.725806…	Frequency of association between two *Macaca fuscata* performing grooming.
Edge_Timestamps	Numeric (integer)	1	Timestamp of the observation (in day number).

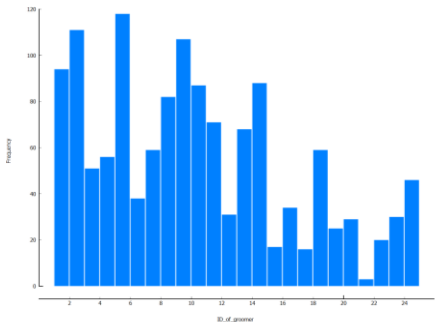

Fig. 1. Frequency distribution of ID_of_groomer.

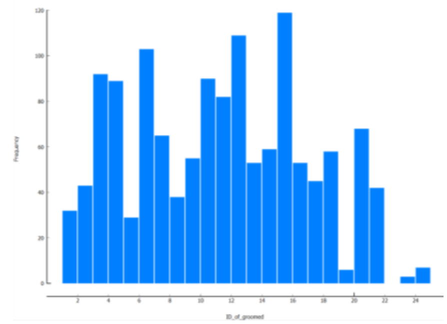

Fig. 2. Frequency distribution of ID_of_groomed.

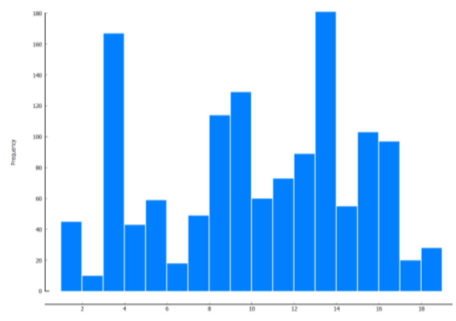

Fig. 3. Frequency distribution of Edge_Timestamps.

grooming or being groomed. Of the data, primate 5 had the largest frequency with a total of 118 instances involving primate 5 as the groomer and primate 21 had the least with a total of 3. Regarding groomed primates, the primate with the largest frequency of being groomed was primate 15 with 119 instances and the primate with the least was primate 22 with 0 instances. Additionally, the "Edge_Timestamps" attributes seem to demonstrate that there were only 19 days of observations. In other words, there are only timestamps up to day 19, implying that there were only 19 days of observations over the original 3-month period this data was obtained.

III. METHODOLOGY

To perform the link analysis, the open graph visualization tool Gephi (version 0.9.2) was utilized. The flowchart displayed in Figure 4 provides a summary of the steps taken to perform the link analysis. First, the data was obtained and imported into Gephi. The data was imported as an adjacency list with a space separator and the character set (charset) of UTF-8. Furthermore, the time representation used were timestamps. Then, the network representation was generated utilizing an undirected graph with auto-scaling, self-loops enabled, and missing nodes allowed to be made. The edge merge strategy used for the network representation was "Sum". Then various global qualities of the network were determined. More specifically, the average degree of each node, the average weighted degree of each node, the network diameter, the graph density, modularity, average clustering coefficient, and average path length were calculated. Each of these qualities were calculated using the "Network overview" features in Gephi with the setting marked to "Undirected graph". For the network diameter, the centralities were not normalized. The modularity was calculated utilizing edge weights and a randomized approach with a resolution of 1.0. The average path length was also run without

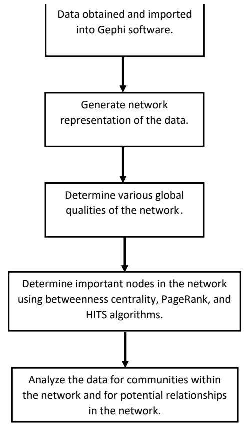

Fig. 4. Flowchart of link analysis methodology.

normalizing the centralities. Then, to determine the important nodes of the network, the betweenness centrality, PageRank, and Hyperlink Induced Topic Search (HITS) algorithms were utilized. The betweenness centrality was calculated for the undirected graph and without normalizing centralities (this was done at the same time as calculating the network diameter since that step is included in determining the diameter). The PageRank algorithm was run on the undirected graph utilizing a probability (p) value of 0.85, an epsilon value of 0.001 as well as utilizing the edge weight. The HITS algorithm was run on the undirected graph utilizing an epsilon value of 0.00001. Of these algorithms, all results for each node were sorted by the highest generated scores and the results were analyzed to determine communities within the network and to determine any potential relationships that may emerge from the analysis.

IV. Results and Discussion

A. Results

The network was generated and can be seen in Figure 5 as an initial basic representation. The network has a total of 514 nodes with 4020 edges. A Fruchterman Reingold representation was also generated for more clear visualization of the network shown in Figure 6. The representation in Figure 6 has generated with the colors of nodes representing the modularity classes and the size of the nodes representing the betweenness centrality of the node. Furthermore, in Figures 7-11, the distribution graphs of each global qualities of the network are shown. The data is also summarized in Table II. The average degree of the network was determined to be 5.21. The average weighted degree of the network was determined to be 15.642. The network diameter is 4 and the graph density was 0.01. Furthermore, the modularity of the network was 0.164. The average clustering coefficient was 0.89 and the average path length of the network was determined to be 2.726. Also, the distribution graphs of the HITS and the PageRank algorithms can be seen in Figures 12 and 13 (betweenness centrality algorithm results are shown in Figure 9 also). The top ten highest scoring nodes for the betweenness centrality, HITs, and PageRank algorithms are also shown in Tables 3-5 and sorted from highest to lowest scoring for each table.

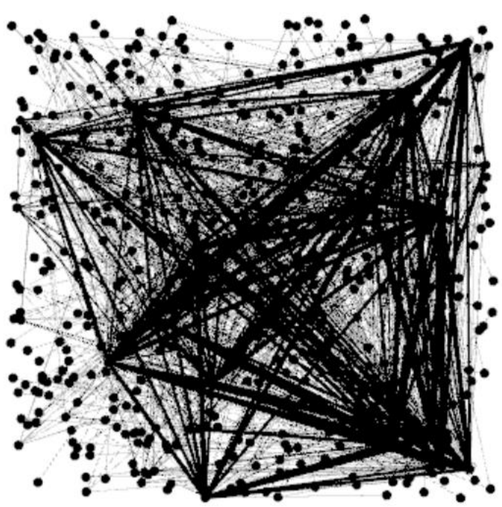

Fig. 5. Initial basic representation of the network.

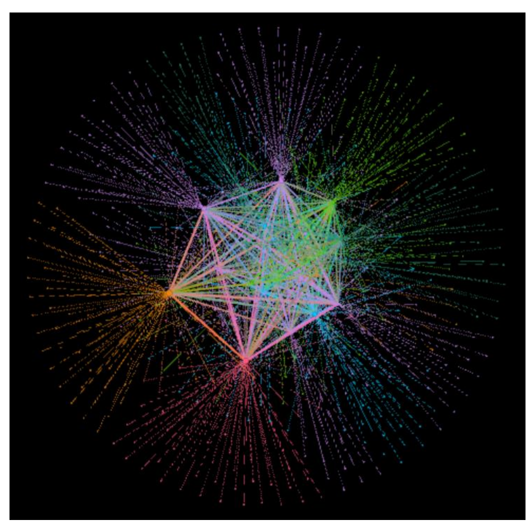

Fig. 6. Fruchterman Reingold representation of the network.

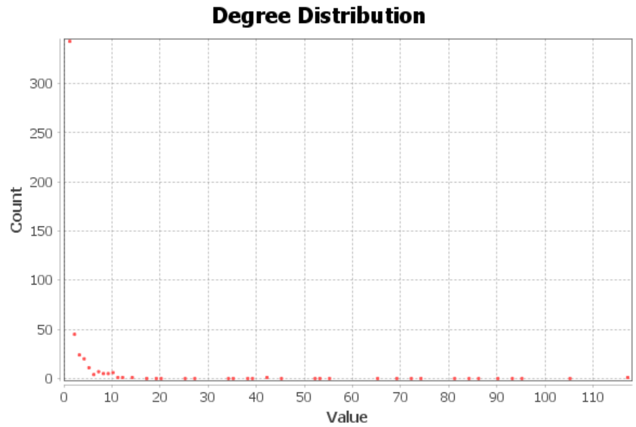

Fig. 7. Average degree distribution of the nodes within the network.

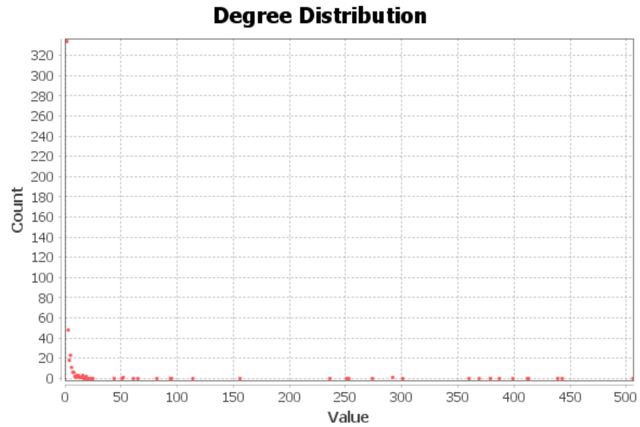

Fig. 8. Average weighted degree distribution of the nodes within the network.

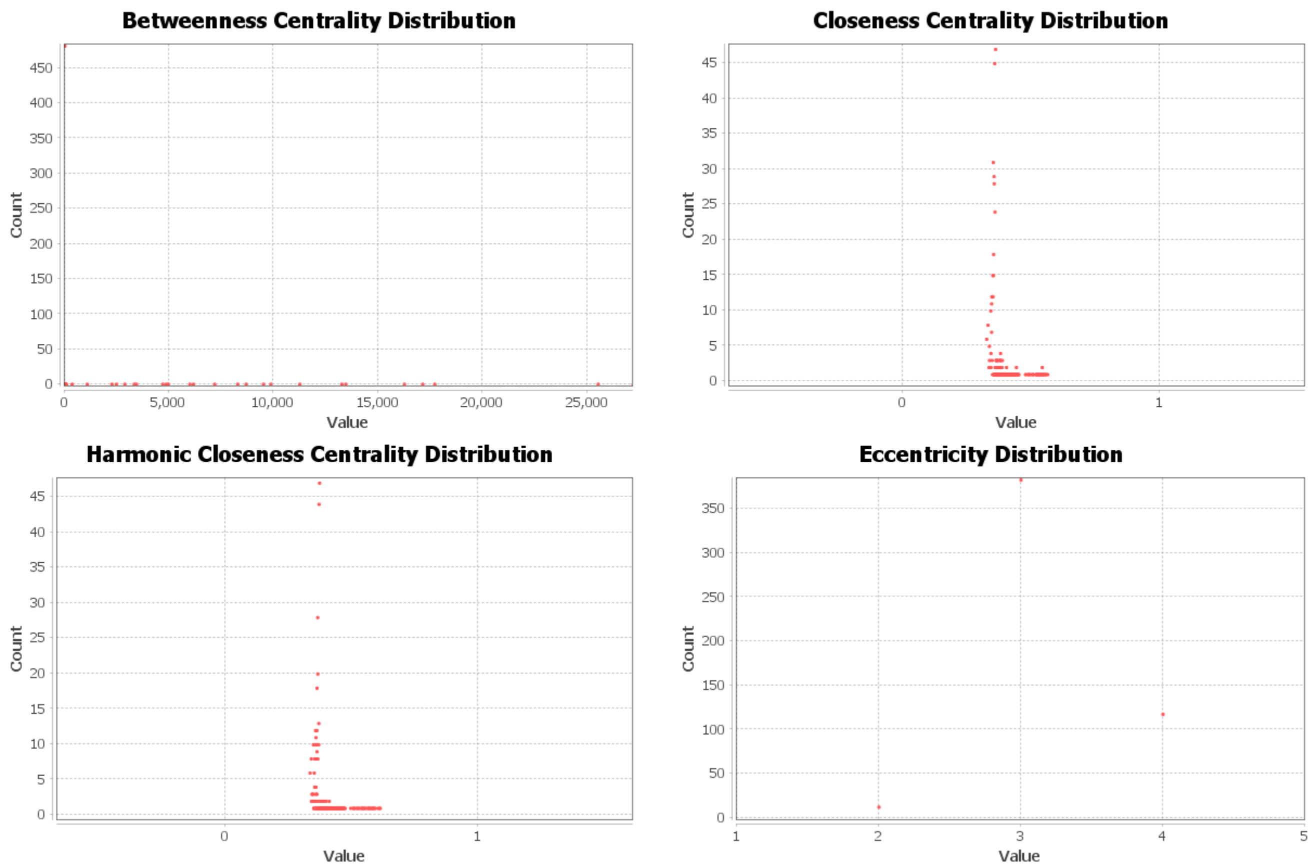

Figure 9. Resulting distribution graphs of the algorithms utilized for calculating the network diameter. The betweenness centrality, closeness centrality, harmonic closeness centrality, and eccentricity of the nodes were utilized to determine the overall network diameter obtained using algorithm referenced [4].

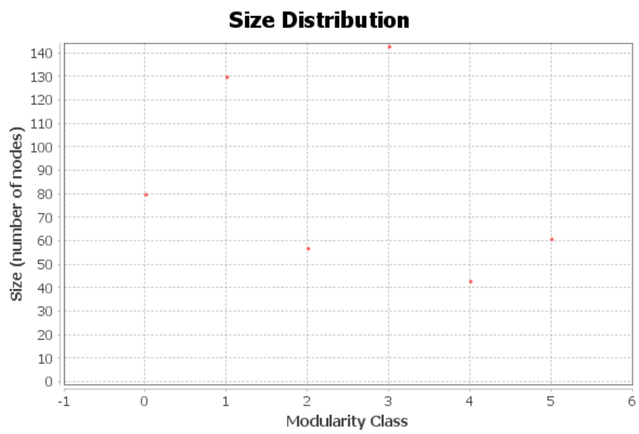

Fig. 10. Modularity class size distribution of the network using algorithms referenced [5 ,6].

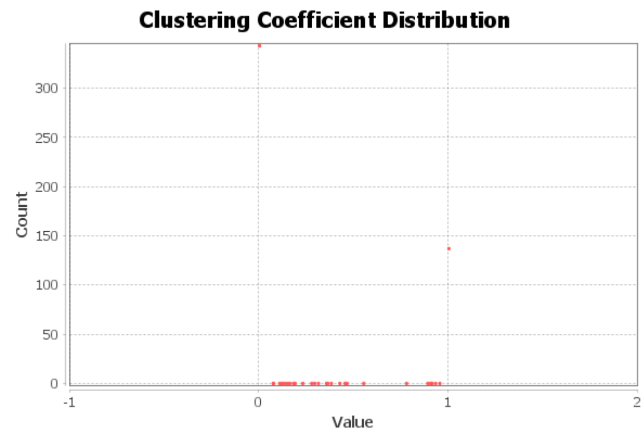

Fig. 11. Average clustering coefficient distribution of the network using algorithm referenced [7].

TABLE II. SUMMARY OF THE NETWORK QUALITIES

Quality	Average Degree	Average Weighted Degree	Network Diameter	Graph Density	Modularity	Average Clustering Coefficient	Average Path Length
Network Value	5.21	15.642	4	0.01	0.164	0.89	2.726

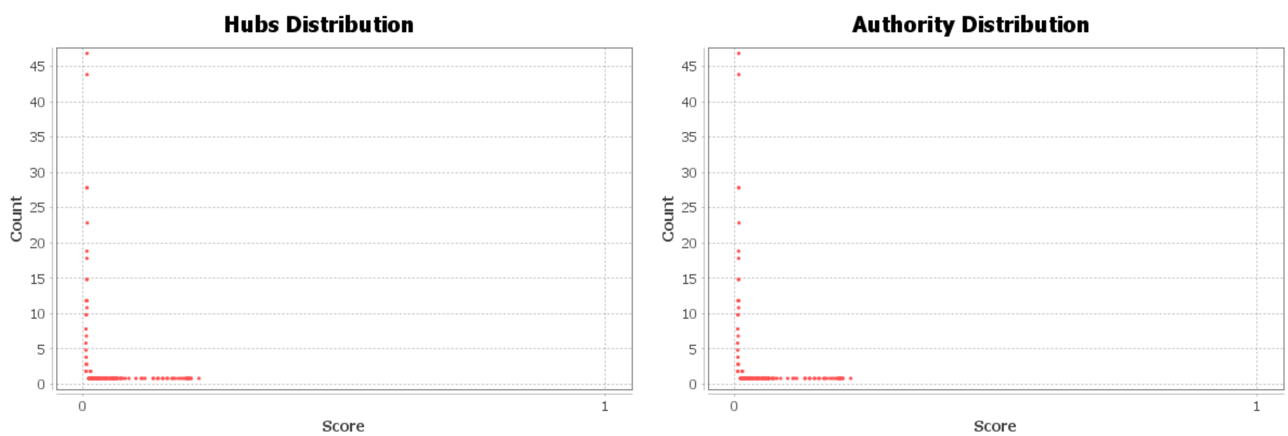

Fig. 12. Resulting Hubs and Authority distribution graphs of the HITS algorithm (algorithm referenced) [8].

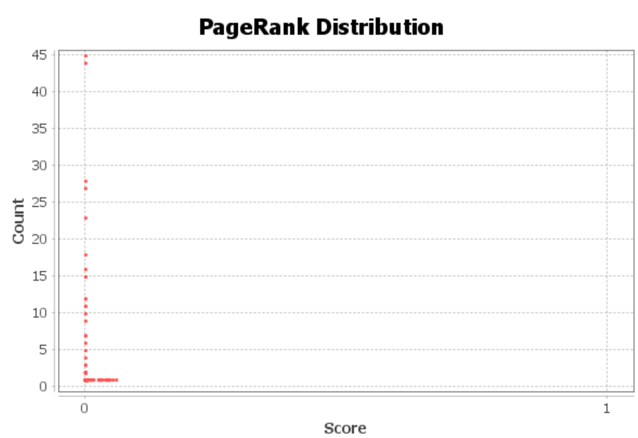

Fig. 13. Resulting PageRank distribution graphs of the PageRank algorithm (algorithm referenced) [9].

TABLE III. TOP TEN NODES WITH HIGHEST BETWEENESS CENTRALITY SCORES IN THE NETWORK FROM THE BETWEENESS CENTRALITY ALGORITHM.

Node	Degree	Weighted Degree	Betweeness Centrality	Modularity Class	Clustering Coefficient
9	117	505	27166.99477	4	0.072263868
5	117	442	25500.18904	2	0.073163418
14	95	378	17684.00853	1	0.110958591
2	105	386	17110.73554	0	0.106043956
10	93	411	16232.28193	1	0.124462494
1	90	359	13427.74124	0	0.12206333
8	86	398	13238.65236	3	0.137535014
13	84	438	11232.66019	5	0.150161622
4	72	300	9846.343343	1	0.179074447
11	81	368	9495.905225	1	0.160443038

TABLE IV. TOP TEN NODES WITH HIGHEST HUBNESS SCORES IN THE NETWORK FROM THE HITS ALGORITHM.

Node	Degree	Weighted Degree	Authority	Hubness	Modularity Class	Clustering Coefficient
2	105	386	0.220792	0.220792	0	0.10604396
10	93	411	0.205911	0.20591	1	0.12446249
5	117	442	0.205354	0.205354	2	0.07316342
9	117	505	0.202731	0.202731	4	0.07226387
13	84	438	0.201303	0.201303	5	0.15016162
11	81	368	0.200477	0.200477	1	0.16044304
14	95	378	0.198971	0.198971	1	0.11095859
8	86	398	0.198014	0.198014	3	0.13753501
1	90	359	0.196944	0.196944	0	0.12206333
7	74	291	0.194705	0.194705	3	0.18569254

TABLE V. TOP TEN NODES WITH HIGHEST PAGERANK SCORES IN THE NETWORK FROM THE PAGERANK ALGORITHM.

Node	Degree	Weighted Degree	Modularity Class	PageRank	Clustering Coefficient
9	117	505	4	0.059221	0.07226387
5	117	442	2	0.0526143	0.07316342
13	84	438	5	0.0460746	0.15016162
10	93	411	1	0.0457198	0.12446249
2	105	386	0	0.0444529	0.10604396
14	95	378	1	0.0437443	0.11095859
8	86	398	3	0.0418068	0.13753501
3	69	412	1	0.0416093	0.18656716
1	90	359	0	0.0396843	0.12206333
11	81	368	1	0.0383555	0.16044304

Table III is a list of the top ten nodes with the highest betweenness centrality scores in the network filtered from highest to lowest. The highest betweenness centrality in the network was 27166.99477 and the lowest was 9495.905225. The degree of the nodes ranges from 72 to 117 with weighted degrees ranging from 359 to 505. Of the nodes, 2 of them belong to modularity class 0, 4 of them belong to modularity class 1, and the rest of the nodes in the table are distributed equally (1 node per modularity class 2, 3, 4, and 5). Furthermore, the clustering coefficients for the nodes ranged from 0.072263868 to 0.179074447. The node with the highest betweenness centrality score was node 9 with a value of 27166.99477. Node 9 also had a degree of 117, weighted degree of 505, clustering coefficient of 0.072263868, and the node belonged to modularity class 4.

Table IV is a list of the top ten nodes with the highest hubness scores in the network that is filtered from highest to lowest. The highest hubness score in the network was 0.220792 and the lowest of the table was 0.194705. The degree of the nodes ranged from 74 to 117 and the weighted degrees range from 291 to 505. Of the nodes, 2 of them belonged to modularity class 0, 3 of them belonged to modularity class 1, 1 node belonged to modularity class 2, 2 nodes belonged to modularity class 3, 1 node belonged to modularity class 4, and 1 node belonged to modularity class 5. Further, the clustering coefficients ranged from 0.07226387 to 0.18569254. The node with the highest hubness score was node 2 with a value of 0.220792. Node 2 had a degree of 105, weighted degree of 386, clustering coefficient of 0.10604396, and it belonged to modularity class 0.

Table V is a list of the top ten nodes with the highest PageRank scores in the network filtered from highest scores to the lowest. The highest PageRank score in the network was 0.059221 and the lowest was 0.0383555. The degree of the nodes ranged from 69 to 117 and the weighted degree of the nodes ranged from 359 to 505. Of the nodes, 2 of them belong to modularity class 0, 4 of them belong to modularity class 1, and the rest of the nodes in the table are distributed equally (1 node per modularity class 2, 3, 4, and 5). Also, the clustering coefficients ranged from 0.7226387 to 0.18656716. The node with the highest PageRank score was node 9, with a score of 0.059221.

B. Discussion

Overall, this link analysis was relatively insightful on the network data. The average degree of the nodes in the graph was 5.21, implying that each node had an average of 5.21 vertices connecting to ~5 other nodes. The average weighted degree was 15.642, which implies that on average the sum of the total degree weights for each node was 15.642. Thus, between these two traits, it can be implied that the average degree weight per degree is approximately 3.12. Furthermore, the network diameter is 4, which implies that the shortest distance between each of the most distant nodes in the network are only 4 nodes apart. The graph density is 0.01 implying that there are only 1% of the possible number of edges in the graph. This low density indicates that the network is a loosely knit network that may not have the strongest of connections between the nodes. The modularity of the network is 0.164 which implies that the community division is slightly divided, but it is not strongly divided. The average clustering coefficient was 0.89 which implies that, on average, the neighborhood of a given node (surrounding nodes of a node) is relatively well connected. Finally, the average path length was 2.726 which indicates that on average there are ~3 steps required for the shortest paths of all possible pairs of nodes within the network.

The network retains a diameter of 4 even though there are 514 nodes total in the network. Although this number of nodes is not necessarily that large in retrospect, the small diameter in comparison to the number of nodes is an indication that the network may be exhibit the small-world phenomenon. The small-world phenomenon implies that any two individual nodes are likely connected via a short sequence of intermediate nodes [10]. This is further corroborated by the high average clustering coefficient (0.89) and the small average path length (2.726), as small-world networks generally have high clustering coefficients and low average path lengths [11]. This would make sense for the data especially given the fact that there are only 25 primates in the whole study that are observed. Consequentially, it is more than likely that a primate would groom or be groomed by any one of the other 24 primates in the study.

Of the nodes in the network, it was determined that, per the betweenness centrality algorithm, that nodes 9, 5, 14, and 2 are the most important nodes of the network. This is largely given to the fact that these nodes had the highest betweenness centrality scores (27166.99477, 25500.18904, 17684.00853, and 17110.73554 respectively). With respect to the HITs algorithm, it was determined that the most important nodes were 2, 10, 5, and 9. This is based on the highest hubness, and authority values generated by the algorithm (0.220792, 0.205911, 0.205354, 0.202731). Finally, with respect to the PageRank algorithm, the most important nodes were determined to be 9, 5, 13, and 10 since the PageRank scores were the highest of the entire network (0.059221, 0.0526143, 0.0460746, 0.0457198). Among these three algorithms, nodes 9, 5, and 2 were frequently determined to be the most important nodes. These nodes have the highest degrees in the entire network (117, 117, and 105 respectively) and some of the highest weighted degrees (505, 442, 386 respectively). Furthermore, node 9, node 5, and node 2 represent primate number 9, 5, and 2, which all had the one of the top three highest groomer interactions in the entire network. Consequentially, it makes sense that these nodes are incredibly important to the network given that many points go through these points. This could also be an indication that these primates play a bigger part in the community of primates observed with respect to grooming. Since these primates mainly performed grooming rather than received it, it would imply that these primates may not have been involved heavily in mutual grooming. Mutual grooming is typically indicative of social bonding and, potentially but less so, mating advantage [12]. Furthermore, it is important to note that per the dataset, all mother/child grooming interactions were not counted in the dataset, thus these higher groomer frequencies could not be attributed to mother/child interactions[2]. Given this information, it could be speculated that these primates were high socially ranking males that performed social bonding for mating advantage.

Per the communities of the network, there were a total of 6 modularity classes determined, indicating there are 6 communities in the given network. However, given the fact that the modularity score of the network was only 0.164, the community separation may not be strongly divided, which means there may be some overlap between the communities. The nodes with the highest degree for each modularity class are as follows: class 0 - node 2 (105), class 1 - node 14 (95), class 2 – node 5 (117), class 3 – node 8 (86), class 4 -node 9 (117), and class 5 – node 13 (84). These nodes with the highest degree in each community also tend to groom rather than be groomed. With the speculation previously discussed with important nodes, those with higher importance in these communities could indicate socially higher-ranking males that perform grooming as a social bonding for mating advantage. Given the fact that the *Macaca fuscata* are polygynandrous, males pair with multiple females and vice versa, it is possible that the nodes belonging to these communities are likely female, especially since they have higher frequency of being groomed rather than grooming within the community [13]. Thus, the communities in the network may actually represent actual small communities of the *Macaca fuscata* in the wild. If this were the case, then the higher-ranking males would be considered the alpha males and the fact that they are performing more grooming than they are receiving it would make a lot of sense. However, if these communities represent actual communities of these primates, then it may be difficult to distinguish specific labels for these communities. Consequentially, the communities' labels would be something like *Macaca fuscata* community 1-6 where the number the community is dependent upon the modularity class.

Furthermore, with respect to the low modularity scores, this may be due how the lower ranking males of the *Macaca fuscata* interfering with the social hierarchy. Non-alpha males of the *Macaca fuscata* perform one of two different forms of mating to elude the social hierarchy: “sneaking off” and “cross-boundary mating” [14]. In cross-boundary mating, the male leaves the group it currently resides in to find a female in a neighboring group and return after the mating is done [14]. This cross-boundary mating could be the reason that the modularity score is low, and the community separation is not that strong. Since the lower ranking males leave these communities and contribute to another it makes sense that the separation would be more muddled.

V. Conclusions

Throughout the course of this report, a link analysis was utilized to determine potential relationships in *Macaca fuscata* grooming interactions. It was determined that the grooming interactions of the given dataset may be heavily representative of alpha males attempting to mate with females in small communities of the primates. This does not mean to exclude the possibility that grooming for sake of strengthening bonds is being performed, rather many of the interactions may insinuate that there are more potential instances of mating from said grooming actions. To fully elucidate this relationship in the data, more observations of these interactions should be obtained and further analysis upon the information from these gatherings should be performed. This additional data may provide further insight on the relationship that is suggested by these conclusions and may also demonstrate new intuition upon the grooming social interaction.

References

[1] D. Krueger, L. Jin, “Social Grooming in Primates,” Reed.edu, 2008. [Online]. Available: https://www.reed.edu/biology/professors/srenn/pages/teaching/web_2008/dklj_site_final/adaptive.html. [Accessed: June 20, 2021].

[2] R. H. Griffin, C. L. Nunn, “Community structure and the spread of infectious disease in primate social networks," Evolutionary Ecology, vol. 26, no. 4, pp. 779-800, 2012. [Online]. Available: https://dash.harvard.edu/bitstream/handle/1/8715733/Griffin_CommunityStructure.pdf?sequence=1&isAllowed=y [Accessed: June 20, 2021].

[3] R.A. Rossi, N. K. Ahmed, “Mammalia-Primate-Association,” The Network Data Repository with Interactive Graph Analytics and Visulaizations, 2015. [Dataset]. Available: http://networkrepository.com/mammalia-primate-association.php. [Accessed: June 2021].

[4] U. Brandes, “A Faster Algorithm for Betweenness Centrality,” Journal of Mathematical Sociology, vol. 25, no. 2, pp. 163-177, 2001. [Algorithm].

[5] V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, “Fast unfolding of communities in large networks,” Journal of Statistical Mechanics: Theory and Experiment, vol. 10, pp. 1000, 2008. [Algorithm].

[6] R. Lambiotte, J. C. Delvenne, M. Barahona, “Laplacian Dynamics and Multiscale Modular Structure in Networks,” IEEE Transactions on Network Science and Engineering, vol 1, no. 2, pp. 76-90, 2015. [Algorithm].

[7] M. Latapy, “*Main-memory Triangle Computations for Very Large (Sparse (Power-Law)) Graphs”*, Theoretical Computer Science (TCS), vol. 407 , no. (1-3), pp. 458-473. 2008. [Algorithm].

[8] J. M. Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, Journal of the ACM, vol. 46, no. 5, pp. 604–632. 1999. [Algorithm].

[9] S. Brin, L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Seventh International Conference World Wide Web Conference (WWW1998), April 14-18, 1998, Brisbane, Australia. [Algorithm].

[10] J. Kleinberg, “The Small-World Phenomenon: An Algorithmic Perspective,” Cornell.edu, 1999. [Online]. Available: https://www.cs.cornell.edu/home/kleinber/swn.d/swn.html. [Accessed: 24 June, 2021].

[11] J. Golbeck, “Network Structure and Measures,” Analyzing the Social Web, pp. 25-44, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124055315000031. [Accessed: June 26, 2021].

[12] S. Linville, “Monkey Bonding Through Grooming,” Indiana Public Media, August 2, 2019. [Online]. Available: https://indianapublicmedia.org/amomentofscience/monkey-bonding-grooming.php. [Accessed: June 26, 2021].

[13] B. Hardman, “Macaca fuscata,” Animal Diversity Web. [Online]. Available: https://animaldiversity.org/accounts/Macaca_fuscata/. [Accessed: June 26, 2021].

[14] Y. Otani, “Brains over brawn: Mating strategies in Japanese macaques,” Research OUTREACH, 2020. [Online]. Available: https://researchoutreach.org/articles/brains-over-brawn-mating-strategies-japanese-macaques/. [Accessed: June 27, 2021].