Distribution Agreement

Yizhou (Joyce) Wang March 27, 2023

Temporal Resolution vs. Accuracy of Rhesus Macaques' Social Network Based on RFID Tracking Data

by

Yizhou (Joyce) Wang

Dr. Gordon J. Berman

Adviser

Dr. Benjamin Wilson

Adviser

Quantitative Theory and Methods

Dr. Gordon J. Berman

Adviser

Dr. Benjamin Wilson

Adviser

Dr. Michal Arbilly

Committee Member

Dr. Weihua An

Committee Member

2023

Temporal Resolution vs. Accuracy of Rhesus Macaques' Social Network Based on RFID Tracking Data

By

Yizhou (Joyce) Wang

Dr. Gordon J. Berman

Adviser

Dr. Benjamin Wilson

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2023

Abstract

Temporal Resolution vs. Accuracy of Rhesus Macaques’ Social Network Based on RFID Tracking Data

By Yizhou (Joyce) Wang

The adaptation of the tracking device enables primate studies to obtain more comprehensive and longitudinal behavior data of the animals. Nevertheless, such data often comes with enormous computational challenges. In the previous relevant studies, it was discovered that data obtained at a very short timescale is comparable with the actual human observational data (Gelardi et al., 2020). In addition to the computational challenges, few studies focused on using longitudinal tracking data to generate and infer the social network of the rhesus macaque. Therefore, this study is aimed to evaluate the accuracy vs. temporal resolution for generating the social network of Rhesus Macaque with 154 days of data collected from 2021 March 4th to 2021 August 4th. Through investigating the spectral clustering result and the Euclidean distance of the proximity matrices generated from different averaging-over durations and the correlation of them with the actual matrilineal genetic similarities, it was found that averaging-over durations of 14 or 28 days are potentially optimal for generating a stable social network that can represent the entire period between March 4th, 2021 and August 4th, 2021.

Temporal Resolution vs. Accuracy of Rhesus Macaques' Social Network Based on RFID Tracking Data

By

Yizhou (Joyce) Wang

Dr. Gordon J. Berman

Advisers

Dr. Benjamin Wilson

Advisers

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2023

Acknowledgements

I would like to thank Dr. Gordon J. Berman and Dr. Benjamin Wilson for providing me with extensive professional guidance in conducting this research. I would also like to thank Dr. Michal Arbilly and Dr. Weihua An for providing valuable advices to my thesis. In addition, I would like to thank Michael Hess and Kanishk Jain for assisting me with processing the data. I would also like to thank Dr. Kim Wallen for providing the data.

Table of Contents

List of Figures

Introduction

A large proportion of nonhuman primates' behavior is social and such behavior connects tightly with their social structure (Brent et al., 2013). Understanding the social structure of non-human primates plays a critical role in animal behavior research and public health practice, such as understanding the persistence of cooperative behaviors, rate of disease, and information spreading (Brent et al., 2013; Gardner & Luciw, 2008). Having knowledge of accurate long-term animal social network structure also lays the foundation of epidemiological models for detecting animal-borne disease transmission (Keeling & Eames, 2005). Moreover, the network approach of non-human primates' social structure provides significant insight into their group formation and stability mechanism in terms of conflicts, reconciliation, conflict intervention, and kinship structure, and ultimately leads to critical insights into primates' health, behavior, and development (McCowan et al., 2011; Hrolenok et al., 2018). With advancements in technology, animal behavior studies can use positional tracking devices to collect continuous observational data that can overcome the time-cost limitations of traditional observational data collection methods (Hrolenok et al., 2018; Gelardi et al., 2020). However, such continuous tracking data often presents computational challenges due to the extensive amount of data available, while the actual social network of the species may remain stable for a certain continuous period (Gelardi et al., 2020). In addition, only a few studies have been applied to track data in inferring rhesus macaque's social structure and none has focused on using longitudinal tracking data to infer and analyze the social structure of rhesus macaque (Hrolenok et al., 2018). Therefore, this study aims to infer and investigate social networks of Rhesus Macaques generated with different averaging-over durations (which is the length of timeframes of the data used for generating the social network) within a half-year time frame based on RFID tracking device collected data. It also aims to determine the optimal averaging-over duration for generating a relatively representative and stable social network of the sampled rhesus macaque in different short-term timeframes based on their half-year tracking data.

Background

The Significances of Rhesus Macaque in Psychological and Biomedical Studies

Rhesus Macaques (*Macaca Mulatta)*, as the Old World nonhuman primates that are geographically distributed across Asia, are one of the most widely used in research (Lewis & Prongay, 2015). The importance of studying their social structure is underlined by the fact that macaque species share important features with humans in terms of physiology, cognition, and social complexity (NPRC). They are ideal candidates for examining the underlying mechanisms of human temperament and studying psychopathology due to their shared social and biological characteristics (Kalin & Sheltona, 2003). They exhibit cognitive imitation and certain theory of mind capabilities, which indicates their significance in both human and animal cognition research (Subiaul et al., 2004; Drayton & Santos., 2016). In 1953, Japanese researchers discovered the notable cultural behavior of the Japanese Macaques— food washing— where the behavior of washing sweet potatoes before eating spread from one young female macaque to the whole troop (Kawamura, 1959 as cited in Fiore et al., 2020). Such cultural behavior reveals the significance of Macaques' social learning capability and the importance of social learning in their emergence (Fiore et al., 2020). Moreover, a past study adopted Rhesus Macaques as their model organism to study brain network activity patterns and the positive influence of peer presence on the performance of well-learned tasks for the similarity of their social behavior and cognition with humans (Monfardini et al., 2017). Because of the similarities in macaque species' physiology, social behavior, and cognition with humans, many other studies have also adopted them as model organisms to reveal the significance of social learning and behavior. Therefore, understanding their social structure is essential for acquiring behavioral insights from the species.

In addition to their significance in social behavior studies, they are also critical species in modeling infectious disease transmission due to their great similarity to human neurobiology and susceptibility to infectious and metabolic diseases (Gibbs et al., 2007). Macaque species have served as models for more than seventy human infectious diseases investigation including Hepatitis B, Poliovirus, Syphilis, and Dengue, as they can be experimentally infected by the infectious disease (Gardner & Luciw, 2008). A recent study has shown the capability of rhesus macaque as the model organism for investigating Covid-19 disease (Munster et al., 2020). It specifically found similar symptom patterns in rhesus macaque and humans in terms of pulmonary infiltrates, viral loads in the noses and throat, and prolonged rectal shedding (Munster et al., 2020). Therefore, understanding the social structure of rhesus macaques is critical for revealing disease transmission patterns in humans and establishing effective medical measures, given the highly socially spread nature of Covid-19. Overall, the species plays an important role in both biomedical and psychological research for modeling and demonstrating behavior, cognitive mechanisms, and infectious disease transmission. Subsequently, it becomes critical not only to understand the social structure but also to construct accurate social networks with valid data for such species.

Social Structure of Rhesus Macaque

Understanding the social structure and hierarchy of macaque groups can help to reveal the underlying social behavior, social cognition, disease, and information transmission pattern in macaque species, and therefore, is critical to psychology and biomedical studies (Brent et al., 2012; Gardner & Luciw, 2008). Rhesus macaques live in multimale and multifemale social groups (Maestripieri & Hoffman, 2011). These groups have a complex matrilineal social structure resulting from sex-specific patterns of dispersal and philopatry. Males leave the natal group to join a new group at puberty, while females remain with their families in the same group throughout their lives. This gives rise to social structure and dominance hierarchy based around female matrilines (groups of closely related females) as well as a smaller number of adult males, typically at the top of the dominance hierarchy. Within each group, the dominance relationship is generally transitive and the hierarchy structure is linear: alpha male outranks all other individuals in the group and alpha female outranks all of the females and most of the females in the groups; all other individuals have their position in the hierarchy according to their dominance rank (Maestripieri & Hoffman, 2011). Throughout their lives, females maintain stable dominant rank whereas males go through losing their rank when they left the group at puberty, acquiring new rank when joining a new group, and eventually work their rank up by forming alliances with powerful males and females in the group (Maestripieri & Hoffman, 2011). Besides individual dominance relationships, matrilines in the group also have dominance relationships where the high-ranking one tends to be the one with a larger population in the group (Ehardt & Bernstein, 1986). The mechanism behind such a hierarchical structure is not genetic, but often social, involving events such as overthrow, maternal intervention, and alliance formation (Ehardt & Bernstein, 1986). Therefore, monkeys' positions within this social structure have profound effects on many aspects of their social life, often dictating access to resources, mating opportunities, and even future reproductive success. Given the importance of these social structures, automated methods and algorithms to analyze and understand macaques' social networks would provide important insights into our understanding of their social behavior.

Proximity Measurements and the Tracking Device Data

Underlying the complex social events that induce the formation of Rhesus Macaques' hierarchical social structure is their agonistic social behavior and affiliative social behavior. The amount of time spent in close proximity is one of the key measures for measuring the strength of social bonds, or the affiliation between individuals (Maestripieri & Hoffman, 2011). Proximity indicates the individuals are in close enough distance to receive assistance or provide service, such as grooming, in an appropriate time frame (Maddali et al., 2014). Therefore, many animal studies collect observational data of proximity to uncover the social structure of macaques and ultimately evaluate the aforementioned biomedical and psychological significance (Gelardi et al., 2020; Brent et al., 2012).

This study uses data collected by wearable RFID devices, which record high-resolution 3D positional data using radio frequency tags that are sensed and triangulated by RFID sensors positioned around the compound. Compared to time-limited observational data collection, the tracking device allows researchers to continuously sample positional data throughout the desired period (Gelardi et al., 2020). However, the high-resolution GPS data poses significant computational challenges as it samples data at short intervals from seconds to seconds. The social structure of Rhesus Macaques remains relatively stable during short-term sampling, and past research suggests that the aggregated network obtained with the tracking device remains stable at a short time scale and is comparable to the network generated from observational data (Gelardi et al., 2020). To date, only a few studies have used position-tracking devices to infer the social structure of Rhesus Macaques, and none of these studies have focused on using longitudinal positional tracking data to infer and analyze the dynamic social structure of Rhesus Macaques over a long-term period (Hrolenok et al., 2018). Therefore, this study will use the longitudinal tracking device data provided by the National Primate Research Center to infer and visualize the social network of Rhesus Macaques over a long-term period. Ultimately, the study will also determine the optimal averaging-over duration for generating the social network of Rhesus Macaques in different time frames. To clearly define the averaging-over duration, it is referring to the length of the timeframe of the data (in which the unit is days) that is used for generating the social network.

Weighted Adjacency Matrix and Similarity Matrix

To investigate the tracking device data, it is necessary to transform the position data to the distance matrix and the adjacency matrix. Given a set of data points $X \in \mathfrak{R}^{p \times n}$, where n is the number of samples and p is the dimensionality of the data, the distance matrix $D = (d_{ij}) \in \mathfrak{R}^{n \times n}$ is a square matrix that shows the distance between each pair of nodes in the graph. An adjacency matrix is a square matrix that allows the representation of a graph with a square matrix $A = (a_{ij}) \in \mathfrak{R}^{n \times n}$ where each element a_{ij} represents the edge attributes of the edge between vertices i and j (Busato & Bombieri, 2017). The adjacency matrix of a weighted graph stores the weights of the edges between each node. If the graph's edges are undirected, the matrix will be symmetric along the diagonal. Moreover, the similarity matrix $S = (s_{ij}) \in \mathfrak{R}^{n \times n}$ is a square matrix that has each entry s_{ij} storing the similarity between data points x_i and x_j.

Laplacian Matrix

The main components for spectral clustering are graph Laplacian matrices, which have been developed into various forms in the past literature (Luxburg, 2007). In between all different kinds of Laplacian Matrices, there are two variants that are widely used: the unnormalized graph Laplacian matrix and the normalized graph Laplacian matrix. Since we are using the unnormalized graph Laplacian Matrix in this study, we will define it in the following. The unnormalized graph Laplacian matrix is defined as

$$L = D - W,$$

where L stands for the unnormalized Laplacian matrix that is symmetric and positive semi-definite (i.e., all entries in the matrix are larger than or equal to 0), D represents the $n \times n$ diagonal matrix that has the degree of each vertex in the graph, W represents the weighted adjacency matrix of the graph. Following are the mathematical notation and definition for the aforementioned matrices:

For an undirected graph with non-negative weighted edges, the weighted adjacency matrix W,

$$W = (w_{ij}) \in \Re^{n\times n},\ where\ i,\ j = 1,\ \dots,\ n \text{, and}$$

1. If $w_{ij} = 0$, the vertices v_i and v_j are not connected by an edge.
2. If the graph is undirected, $w_{ij} = w_{ji}$.

The degree matrix $D = (d_{ij}) \in \Re^{n\times n}$, the degree of a vertex x_i in X is defined as

$$d_{ii} = \sum_{j=1}^{n} w_{ij}$$

Therefore, by definition, d_{ii} sums over all vertices adjacent to v_i that are connected by edges with weight nonzero. Given the above definition and properties of matrices D and W, the Laplacian Matrix has the following properties:

$$L_{ij} = \begin{cases} -w_{ij} & \text{if } i \sim j \\ w_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Therefore, by definition, as L is an $n \times n$ symmetric and positive semi-definite matrix, it will have *n* non-negative, real-valued eigenvalues. In addition to the purpose of distinguishing the connected components in the graph, the Laplacian Matrix *L* measures the extent that a graph differs from one vertex to another vertex.

Spectral Clustering

Spectral clustering, as one of the most popular modern clustering algorithms, can be efficiently used to identify groups of similarity in the data (Luxburg, 2007). It outperforms the traditional K-means clustering as it overcomes two major issues in the method: the assumption of the spherical shape of the cluster, and the difficulty of determining the cluster centroid (Rodriguez et al., 2016). Specifically, with the similarity matrix, the spectral clustering algorithm will help to identify the partition of the network graph with the edges between different groups that have very low weights and thus, meaning the nodes within one group are very dissimilar from the other one and the edges within one group have higher weights which indicate the member intragroup are very similar (Luxburg, 2007). The general steps of spectral clustering are as follows (Luxburg, 2007):

1. Construct the Similarity Matrix *S* from graph *G* with weighted adjacency matrix *W*.
2. Compute the unnormalized Laplacian *L*.
3. Compute the *k* eigenvectors of *L* corresponding to *k* smallest eigenvalues and denote them as $u_1, u_{2, \dots} u_k$.
4. Let $U \in \mathfrak{R}^{n\times k}$ be the matrix containing $u_1, u_{2, \dots} u_k$ as its column vectors.
5. Cluster the points in each *i*-th row of the matrix *U* with the k-means algorithms into clusters $C_1, C_2, \dots, C_k$.

In summary, the output structure of the spectral clustering is *k* clusters that are respectively grouped together based on their similarity.

K-means Clustering and K-medoids Clustering

As mentioned above, the spectral clustering algorithm uses k-means clustering on the Laplacian matrix *L*'s *k* eigenvectors. The k-means clustering is an iterative data-partitioning algorithm that assigns *n* observations to exactly one of the *k* clusters based on the Euclidean distance of two points (Chang, Huang & Wu, 2009; Lloyd, 1981). The general algorithm of the k-means clustering includes

1. Select the *k* random centroids in data
2. Compute the *n* data point's distance to each centroid using the Euclidean distance.
3. Choose the cluster for the *n* data points by finding the cluster with the centroid that has the minimum distance to reach them.
4. Reinitialize the centroid and repeat steps 2 and 3 until the clusters stop changing.

Since the k-means clustering on the centroid of the cluster, which is not necessarily one of the data points, its clustering results will be greatly affected when there are outlier data (Change, Huang & Wu, 2009). In contrast to k-means clustering, k-medoids clustering chooses the actual data point as the center of the cluster and assigns data to the clusters based on their dissimilarity (e.g., Euclidean distance) with the medoids. By choosing actual data as the center of the cluster, k-medoids clustering is much less sensitive than k-means clustering and provides higher interpretability. The following is the general steps of k-medoids clustering (Park & Jun, 2009),

1. Select the k initial medoids.
2. Compute the distance of each data point to each medoid.
3. Obtain the initial cluster by assigning each data point to the nearest medoid.
4. Find a new medoid of each cluster, which is the data point that minimizes the total distance to other data points.
5. Assign the data point to the nearest medoid and calculate the sum of the distance from all data points to their medoids. If the distance is less than the previous one, use the new medoids as the medoids until the clustering (or the distance to medoids of data points within clusters) stops changing.

Rand Index and Modified Rand Index

To evaluate the clustering result of data without an absolute scheme of clustering, Rand Index provides a proper measure by evaluating the similarity of arbitrary clustering results. Rand Index assumes the clustering is discrete in which all data points are assigned to a specific cluster; it also assumes the cluster is defined as much by the points they contain as by the points they do not contain; and, it assumes equal importance in the clustering determination process (Rand, 1971). With these assumptions, Rand Index aims to measure the similarity between clustering results *Y* and *Y'*, each consisting of *n* points (Rand, 1971). The following formula defines the Rand Index mathematically,

$$Rand\ Index\ (Y,\ Y') \ = \ (a \ + \ b)\ /\ \ {}_{n}C_{2},$$

$where,\ a \ = \ number\ of\ pairs\ of\ elements\ in\ Y\ that\ are\ in\ the\ same\ subsets\ in\ Y'$

$b \ = \ number\ of\ pairs\ of\ elements\ in\ Y\ that\ are\ in\ the\ different\ subsets\ in\ Y'$

${}_{n}C_{2} \ = \ n(n \ - \ 1)/2,\ number\ of\ total\ possible\ pairs$

Therefore, *Rand Index (Y, Y')* represents the similarity between the two clustering results *Y* and *Y'*. Rand Index will be one when the two clustering results are completely the same and will be zero when the clustering results have no similarities. The similarity essentially represents the proportion of pairs of elements whose relationship is the same in both clustering results (Rand, 1971).

Since the study will have data that have a changing size due to the tracking devices' technical issues, Modified Rand Index will be used to evaluate the clustering. The Modified Rand Index accommodates the Rand Index computation for two partitions *U* and *U'* with different sets of units (Cugmas & Ferligoj, 2018). It is defined as the ratio between the number of all possible pairs of units placed in the same or different clusters in both *U* and *U'* and the number of all possible pairs of units. Given *U* has *M* points, *U'* has *N* points, and *M* < *N,* the mathematical formula is defined as below,

$$Modified\ Rand\ Index(U,\ U') \ = \ (c \ + \ d)\ /\ \ {}_{N}C_{2}$$

where, *c = the number of pairs of elements in U that are in the same subsets in U',*

d = the number of pairs of elements in U that are in the same subsets in U',

$_{N}C_{2} \;=\; N(N\,-\,1)/2,\ the\ number\ of\ total\ possible\ pairs.$

The modified rand index will be equal to the rand index when the size of U and U', or *M* and *N*, are the same.

Euclidean Norm

In addition to comparing the clustering results, one can also use the Euclidean norm to evaluate the distance between different matrices by measuring the distances between the eigenvectors of the two matrices. The Euclidean norm of two vectors measures the Euclidean distance between the two input vectors. Given vectors $\vec{v}\ and\ \vec{u}\ \in R^{n}$, the distance between the two vectors is

$$d(\vec{u},\, \vec{v}) \;=\; ||\vec{u} \;-\; \vec{v}\,|| \;=\; \sqrt{(u_{1} - v_{1})^{2} \;+\; (u_{2} - v_{2})^{2} \ldots \;+\; (u_{n} - v_{n})^{2}}\ .$$

The result of the Euclidean norm will be a positive distance value.

Methods

Data Collection

The data of the study is provided by the National Primate Research Center. The data were collected using the RFID tag collar, which records 3D positional data at a rate of around 30 Hz. Each collar has four tags that simultaneously record the positional data. The tracking system operates continuously from 2021 March 4th to 2021 August 4th for recording the macaques wearing the collar tracking device. Therefore, this study will use the 154 days of data collected to generate the social network of 26 macaques and investigate the optimal averaging-over duration among 1 day, 7 days, 14 days, 21 days, …, and 140 days for generating a stable network that can represent the short-term period. The averaging-over duration indicates the length of the sampled days for generating the social network.

Preliminary Processing and Correction of the Data

To reduce the computation difficulty, the data is sampled at a 10-hertz rate for the proceeding analysis. Since the tracking device occasionally experiences technical complexities, such as recording out-of-range data, and rapid jumps in a limited time, the data are processed and filtered with the following steps to ensure accuracy.

1. Removed the data that is out of the range of the tracking field.
2. Removed the tag data for one tracking device that has a distance to the median of the four tags on the collar greater than 50cm.

The final position data for each macaque throughout the collection period is the median of each macaque's corrected collar data (i.e. the median of the four tags on each collar).

Proximity Calculation

With the 3D position data of each monkey, we are able to calculate the distance between each dyad of macaques across each sampling time point during the data collection. The distance is calculated based on the 3D Pythagorean theorem. For macaque A with positional data $\{x_a, y_{a,} z_a\}$ at the time point *t*, and macaque B with positional data $\{x_b, y_{b,} z_b\}$ at the same time point *t,* the distance d_t *(A, B)* between them at the time point *t* is defined as

$$d_t(A, B) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2} .$$

Following Rhesus Macaques' social behavior demonstrated in the background section, a dyad of macaques that is in close proximity to each other is defined as when their distance is less than 1 meter. Summing up the number of times a dyad is in close proximity provides a proximity score for the dyad, which essentially represents how close a dyad of macaques is. Given macaque *A* and *B* with distance *d(A, B) = {*d_1 *(A, B),..., *d_n*(A, B)}* throughout the time period with *n* time points, the proximity score *Proximity(A, B)* between the two macaques is,

$$Proximity(A, B) \ = \ number\ of\ times\ the\ d_i(A, B) \ < 1\, per\ sample,$$

where $i \in n$.

With the proximity scores, we construct the proximity matrix *P* for all macaques within the desired time frame. Each entry p_{ij} is the proximity score *Proximity(i, j)* between macaque *i* and macaque *j*. The proximity matrix *P* for an averaging-over duration *dur* that is in the time frame of $T = \{t_1, t_2, \ldots, t_{dur}\}$, where t_i is associated with a proximity matrix P_i that has all dyads' proximity score, is calculated by aggregating all P_i associated with *T,*

$$P = \sum_{i=0}^{dur} P_i$$

For different time frames, there can be newcomer macaques and outgoer macaques. Although it will include all macaques that have been recorded in the selected time frame, the size of the proximity matrix *P* varies for different time frames.

To normalize the proximity matrices obtained by different averaging-over durations, we divide the cumulative proximity scores of each dyad by the averaging-over durations which stands for the length of time the data was sampled. Given macaque A and B with proximity *Proximity(A, B)*, their normalized proximity score in duration *dur* days 1 day, …, 140 days is

$$Normalized\ Proximity(A,\ B)\ = p(A,\ B)/(dur\ *\ 10\ *\ 3600\ *\ 24)\ ,$$

where the denominator is the product of the number of days *dur* the data being collected and the number of samples per day ($10\ *\ 3600\ *\ 24$).

Spectral Clustering

Selecting The Optimal Number of Clusters k

To evaluate the difference between the networks generated by different averaging-over durations on a normalized basis, we used spectral clustering to identify the clusters from the different proximity matrices. The number of clusters *k* is determined by comparing the rand indexes of the clustering results generated by *k = 1, 2, 3, …, 15* for every window throughout the duration from 1 day to 10 days. For example, for averaging-over duration = 2-day in the total data collection period from 2021.03.04 to

2021.08.04, time windows in this duration will be 2021.03.04 to 2021.03.05, 2021.03.05 to 2021.03.06, 2021.03.06 to 2021.03.07…

Spectral clustering with $k = 1{:}15$ was performed for all time windows in each averaging-over duration from 1 day to 10 days for examining the optimal k for clustering. Each k is associated with a cross-validated modified rand index (MRI) r_k, which is calculated by taking the average of the modified rand indexes between the clustering result U_k (i.e., the spectral clustering results when clustering number = k) and the clustering results U_i (where $i = \{1, 2, 3, \ldots 15\}$ except k) at each time window. The following defines such comparison formulaically:

Given K = {1,..., 15}, $k, i \in K$ where $i \neq k \; and \; i = \{1, 2, \ldots, 15\}$, each k is associated with a clustering result U_k from spectral clustering, and each i is associated with corresponding clustering result U_i

$$r_k = \sum_{i=1}^{14} MRI(U_k, U_i) \,/\, 14$$

To ensure the clustering results can represent the overall data, we shuffle all clustering results U_i for each i, where we obtain the shuffled clustering result V_i. With the shuffled clustering result V_i, we recompute the cross-validated rand indexes for each k. By doing so, we obtain the γ_k for each k, where it is defined as,

$$\gamma_k = \sum_{i=1}^{14} MRI(U_k, V_i) \,/\, 14$$

Therefore, given all windows across averaging-over durations 1 to 10 days (total number of windows = n), each time window within each averaging-over duration is associated with an r_k and an γ_k . To compare across all time windows at each averaging-over duration, the overall average of cross-validated MRI R_k and R'_k corresponding to each k in 1 to 15 at each averaging-over duration were computed as the following:

$$R_k = \frac{\sum_{i=1}^{n} r_k}{number\ of\ windows\ at\ the\ duration}$$

$$R'_k = \frac{\sum_{i=1}^{n} \gamma_k}{number\ of\ windows\ at\ the\ duration}$$

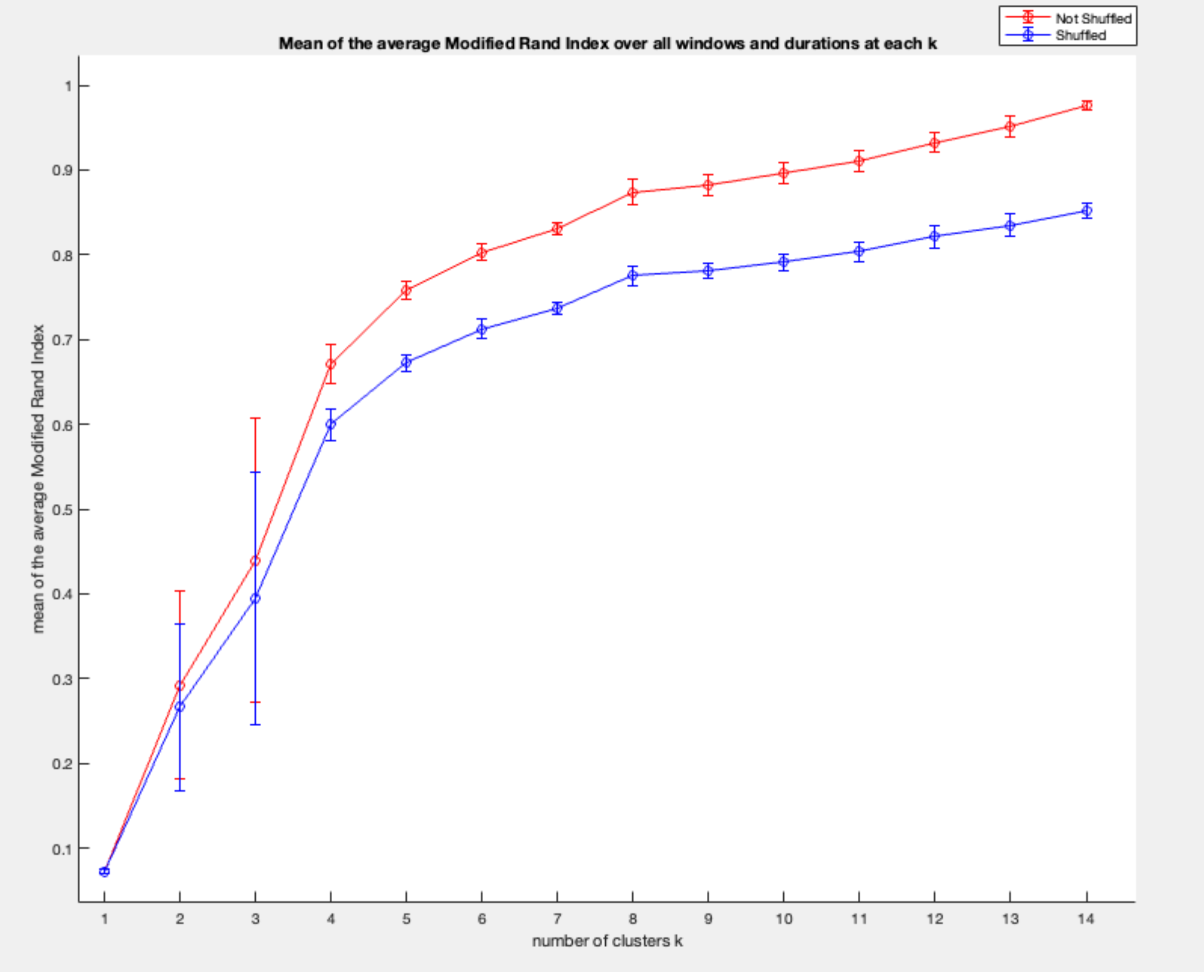
With the average cross-validated MRI for each k at each averaging-over duration, we are able to compute the mean of average cross-validated MRI for each k across all 22 averaging-over durations RR_k and RR'_k as the following:

$$RR_k = \frac{\sum_{i=1}^{n} R_k}{22}$$

$$RR'_k = \frac{\sum_{i=1}^{n} R'_k}{22}$$

Fig. 1. shows the mean of the average of cross-validated rand index of shuffled and non-shuffled clusters RR_k and RR'_k for each $k \in \{1, 2, 3, \ldots, 15\}$ across all averaging-over durations. The elbow point of the increasing curve takes place around k = 6, indicating the clustering result of k = 6 can relatively well represent the clustering results generated from all ks. In other words, the clustering results at the number of clusters $k = 6$ resulting from all different possible windows in the averaging-over durations = 1 to 10 days during the period 2022.03.04 to 2022.08.04 is relatively well-representative or having a relatively high similarity, of macaques' possible social networks that are consisted respectively of one group, two groups, three groups, …, fifteen groups of macaques. Therefore, the proceeding analysis will be based on the clustering result when the number of clusters $k = 6$.

Figure 1. Mean of the average cross-validated Modified Rand Index for k = 1:15 over all windows and averaging-over duration of 1 day to 10 days. The Mean of the average cross-validated MRI for each *k* was computed by finding the mean of the average cross-validated MRI of each k across all time windows at each averaging-over durations throughout all durations (e.g. the mean of the average MRI at *k = 1* represents the average of the 22 average cross-validated MRI across all durations, where each average cross-validated MRI is computed by taking the average of MRI over all windows at each duration). The elbow point of the increasing mean occurs around *k = 6*.



Spectral Clustering With k = 6

For each averaging-over duration in *dur = {1 day, 7 days, 14 days, 21 days …, 140 days},* spectral clustering was performed based on *k = 6* for each time window within the duration. The similarity matrix *S* is used to measure the similarity between the macaques based on the proximity scores. In this case, the proximity matrix *P* was normalized to ensure that each entry p_{ij}, being the normalized proximity score *Normalized Proximity(i, j)* of the macaque i and j, had a value between 0 and 1, which makes it a suitable input for spectral clustering. To transform the similarity matrix into a matrix that is suitable for clustering, the Laplacian matrix *L* was then computed as it can capture the data structures and project them into a low dimensional space so that we can easily identify the *k* clusters. Specifically, the similarity matrix was transformed to the unnormalized Laplacian matrix *L* based on the formula *L = D-S,* where *D* is the degree matrix *d* that was computed based on the following equation

$$d_{ii} = \sum_{j=1}^{n} S_{ij} \qquad .$$

The *k* eigenvectors (*k = 6)* of the Laplacian matrix *L* corresponding to its *k* smallest eigenvalues were then computed and input to a matrix $U \in \Re^{n\times k}$ respectively as its column vectors $u_1, \ldots, u_k$. K-means clustering was then performed in each i-th row of the matrix *U* to form clusters $C_1, C_2, \ldots, C_k$. By doing the above steps of spectral clustering, we are able to group all macaques in the data into 6 clusters based on their pairwise proximities in different time windows that belong to different averaging-over durations and compare the clustering across these time windows.

Selecting the Optimal Window-away Distance

With the optimal k = 6, we are able to compare the windows within each average-over duration by different window-away distances and select the optimal window-away distance for later comparison of the effect of averaging-over durations on the stableness of the clustering results generated by different windows distanced by the optimal window-away distance. By window-away distance, we mean the number of windows between the window we intend to compare. For example, for averaging-over duration

= 2 days and window-away distance = 2 windows, we will compare the window 2021/03/04 ~ 2021/03/05 to window 2021/03/06~2021/03/07. We used the distance between different windows to categorize and compare the windows fairly at each averaging-over duration since the cross-validation method described above will produce an unfair comparison for comparing the rand index of the clustering results of windows that are distanced away by lots of time windows to the rand index of the clustering results of windows that are only distanced away by few time windows. For instance, with the averaging-over duration = 1 day, comparing the modified rand index between the clustering results *(2021/03/04) and (2021/03/05)* to the modified rand index between the clustering results *(2021/03/04) and (2021/08/04)* will be unfair since the later one will be naturally dissimilar, having a much lower MRI than the previous one, to each other as the social structure of macaques developed over the five months period. By the following procedures, we are able to identify the optimal window-away distance, or the distance, that can most fairly compare the clustering results of windows at each averaging-over duration.

For each averaging-over duration *dur*, there are *k* consecutive windows as 1st, 2nd, …, and kth fall into the total data collection period ($k = total\ days\ of\ data\ collection - dur + 1$). Therefore, there are *N* possible distances between different windows for each *dur*, where $N = k - 1$. For each $n \in N$ at each averaging-over duration *dur*, there are *x* window pairs that fall into such a window-away distance bin *n*, where x_i denotes each window pair's clustering results. Therefore, each window-away distance *n* is associated with an average rand index value W_n that indicates how well the clustering results of window pairs falling into the window-away distance *n* bin represent each other. In other words, it represents how stable the clustering results are over all the window pairs that are distanced away by *n* windows for the averaging-over duration *dur*. W_n is defined as the following,

$$W_n = \sum_{i=0}^{x} MRI\,(x_i)\,/\,x\,.$$

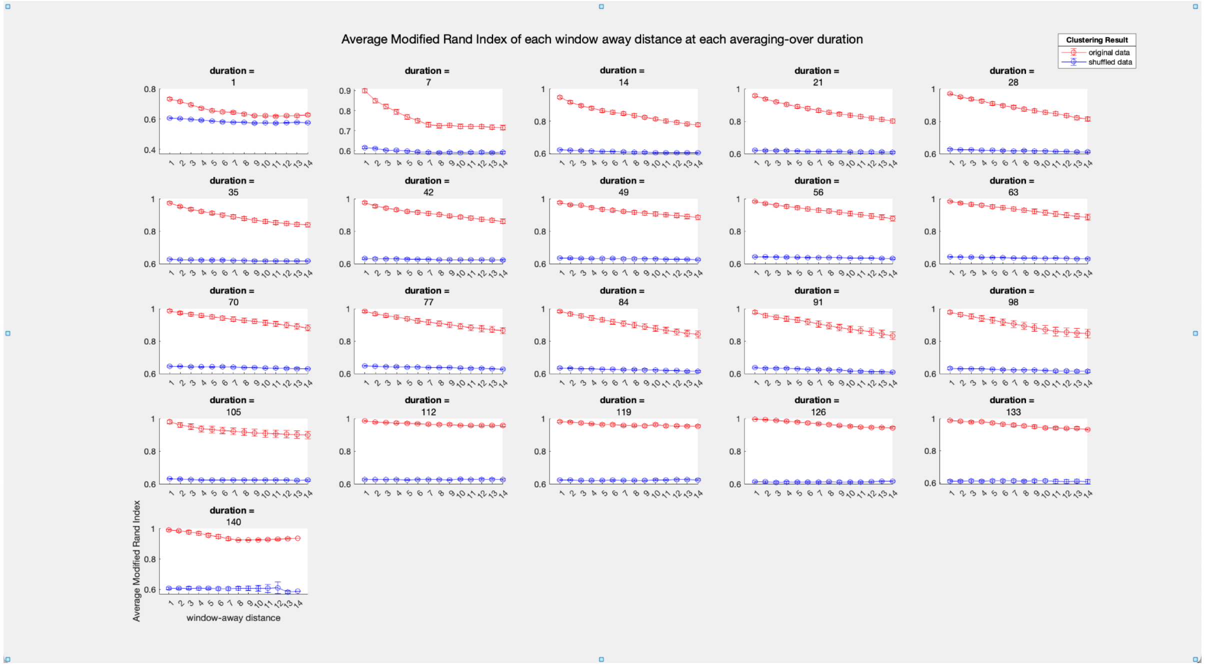
To ensure the clustering results and the W_n calculated is not due to randomness, we shuffled the clustering result for each x_i pair and obtained the new shuffled clustering result v_i. With the shuffled clustering result

v_i, we recompute the average modified rand index ω_n as the following,

$$\omega_n = \sum_{i=0}^{x} MRI\,(v_i)\,/\,x$$

By computing the average rand index W_n and ω_n of each window-away distance for each averaging duration, it is found that the average rand index W_n is highest throughout all different duration when the window-away distance = 1 window as shown in Fig. 2. The shuffled average rand index ω_n shows a significant deviation from the distribution of the actual average rand index W_n, indicating the distribution of actual W_n is not due to randomness. Thus, when window-away distance = 1 window, the clustering results of windows belonging to window pairs that fall into the distance bin can relatively well-represent each other throughout all averaging-over durations. In other words, the social networks generated by the time windows that are distanced by one window, such as 2021/03/04's social network and 2021/03/05's social network for averaging-over duration = 1 day, tend to stay stable across the averaging-over duration. Therefore, comparing the stableness of the clustering results of windows that are distanced by one window at each duration over all averaging-over durations is the most reliable metric to evaluate how different averaging-over durations can influence the stableness of the social networks (proximity matrices) generated by the windows.

Figure 2. Average Modified Rand Index of each window away distance at each averaging-over duration. The average modified rand index at each window-away distance is calculated by taking the average of the MRI of each window pair's clustering results that are distanced by the window-away distance at each case of averaging-over durations (or, the size of each window). Throughout different durations, the windows that are distanced by one-window have clustering results that can most well represent each other.



Selecting the Optimal Averaging Duration

By comparing the average of the modified rand index of the windows across the cases of them distanced by different window-away distances, we were able to identify window-away distance = 1 window provides the metric that can relatively fairly compare the windows, meaning the time windows that are distanced away by one window across all durations have clustering results that are very similar to each other. With this optimal window-away distance, we can compare how the averaging-over duration affected the stability or similarity of clustering results for pairs of windows that were separated by one window at each duration. To compare the averaging-over durations comprehensively, we also include other window-away distances' comparison results. With 154 days of data (collected from 2021.03.04 to 2021.08.04), we are able to generate *k* windows for each averaging-over duration *dur* days, where *k* is defined as the following,

$$k = 154 - dur + 1 .$$

Therefore, there are *k -1 possible* window-away distances for each duration *dur*. Since the maximum averaging-over duration is 140 days, the maximum window away distance to compare across all averaging-over duration is 14 days. Thus, we compare each $W_{n\ at\ duration\ =\ dur}$ associated with each *dur* for each *n,* where n = {1, 2, 3, …, 14}, to examine the variation of the stability of the clustering results generated from windows with different window size (averaging-over duration) when they are separated by *n* windows.

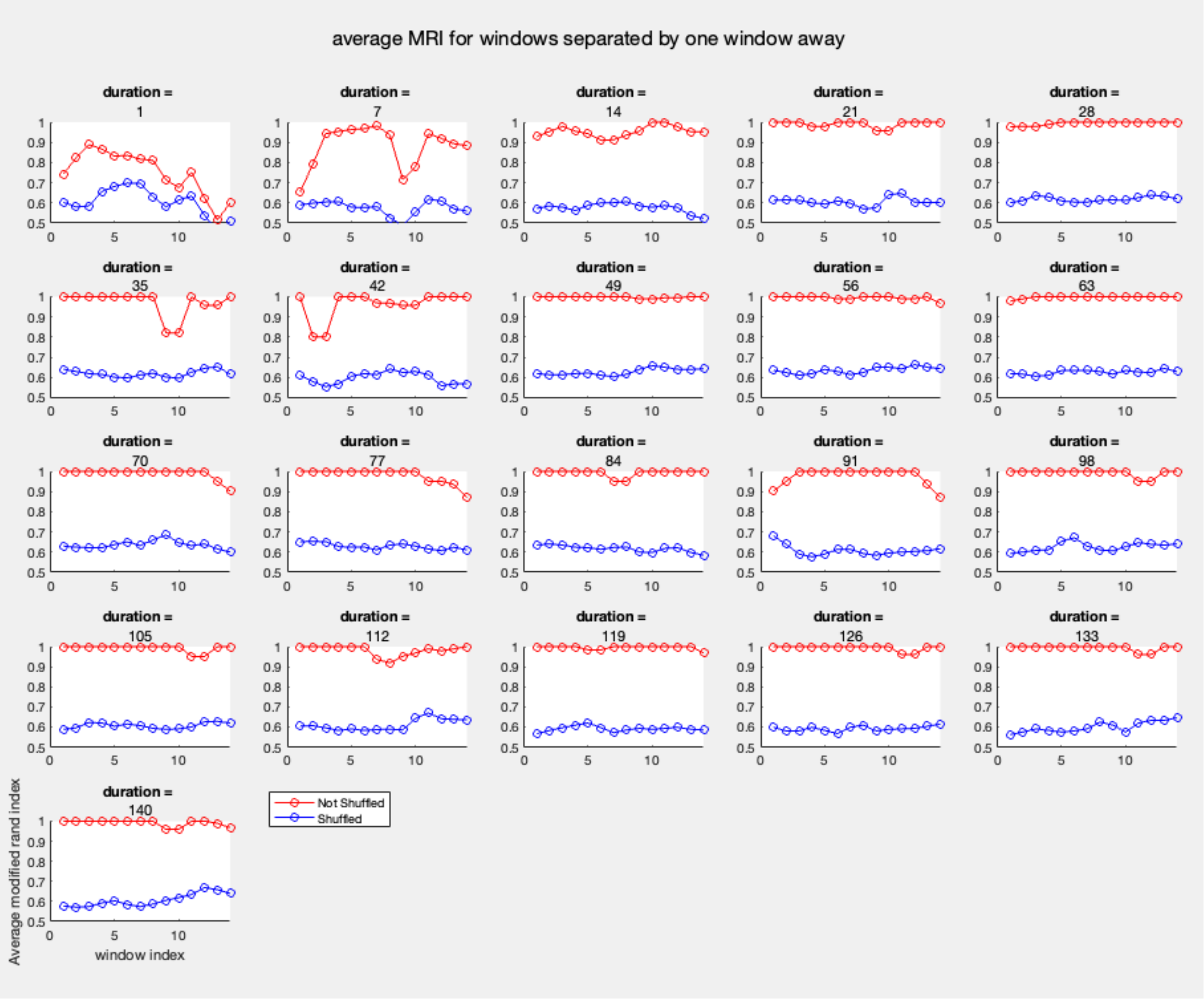
By doing so, we are essentially evaluating how averaging-over duration (window size) can influence the stability or similarity of macaques' social networks generated across the time windows that are each distanced away by *n* windows. For instance, for window-away distance = 1 window, we can compare the average modified rand index of clustering results for averaging-over duration *dur* = 1 day which consisted of $W_{1\ at\ duration\ =\ 1\text{-}day}$ *= average of {MRI(2021/03/04, 2021/03/05), MRI(2021/03/05, 2021/03/06)...}* and the average modified rand index of clustering results for averaging-over duration *dur* = 2-day which consisted of $W_{1\ at\ duration\ =\ 2\text{-}day}$ *= average of {MRI(2021/03/04~2021/03/05, 2021/03/05~2021/03/06), MRI(2021/03/05~2021/03/06, 2021/03/06~2021/03/07)...}.* The aforementioned comparison essentially allows us to identify the optimal averaging-over duration that has the highest average modified rand index of the clustering results across the windows when they are distanced away by all different window-away distances. This optimal averaging-over duration implies an ideal duration that is able to produce clustering results of proximity matrix, or ultimately the social structure, of the macaques that is relatively accurate for representing the social structure over the entire data collection with the minimum amount of data.

We expect the optimal averaging-over duration to have a relatively high average rand index $W_{n\ at\ duration\ =\ dur}$ across all different *n* and especially at *n=1 window-away distance,* indicating its highest stability throughout the time windows it averaging over with. We also expect the optimal averaging-over duration to be at the elbow-point that shows a slowdown of the fluctuation of W_n across all *n.* Since we found window-away distance = 1 window provides us an optimal Modified Rand Index, we will use the one-day window away distance as the priority metric to compare across different averaging duration.

Selecting the Optimal Window Within the Optimal Window-away Distance for Visualization Purposes

To achieve our goal of generating a stable social network that can well-represent the social structure of a short-term time frame, we also locate the optimal time window within the optimal window-away distance for generating stable proximity matrices by calculating the average modified rand index of each window with their one-window-away neighbor windows for each averaging-over duration. Fig. 3. presents the ideal window around window 3 since it has a relatively high average modified rand index across all different durations although there are fluctuations between different averaging-over durations, indicating the third window's social network across all duration have a relatively high similarity with its neighbor windows' social network and thus, can well represent other windows' social network. Therefore, we will use window 3 to visualize the social network and clustering result at the optimal averaging-over duration.

Figure 3. Average Modified Rand Index for each window with their one-window-away neighbor at each averaging-over duration from 1 to 10 days. The average modified rand index of each window is calculated by taking the average of the MRI of the window with its two neighbor windows that is one window distanced away. The optimal average MRI occurs around window 3 for each duration.



Euclidean Distance of Matrices

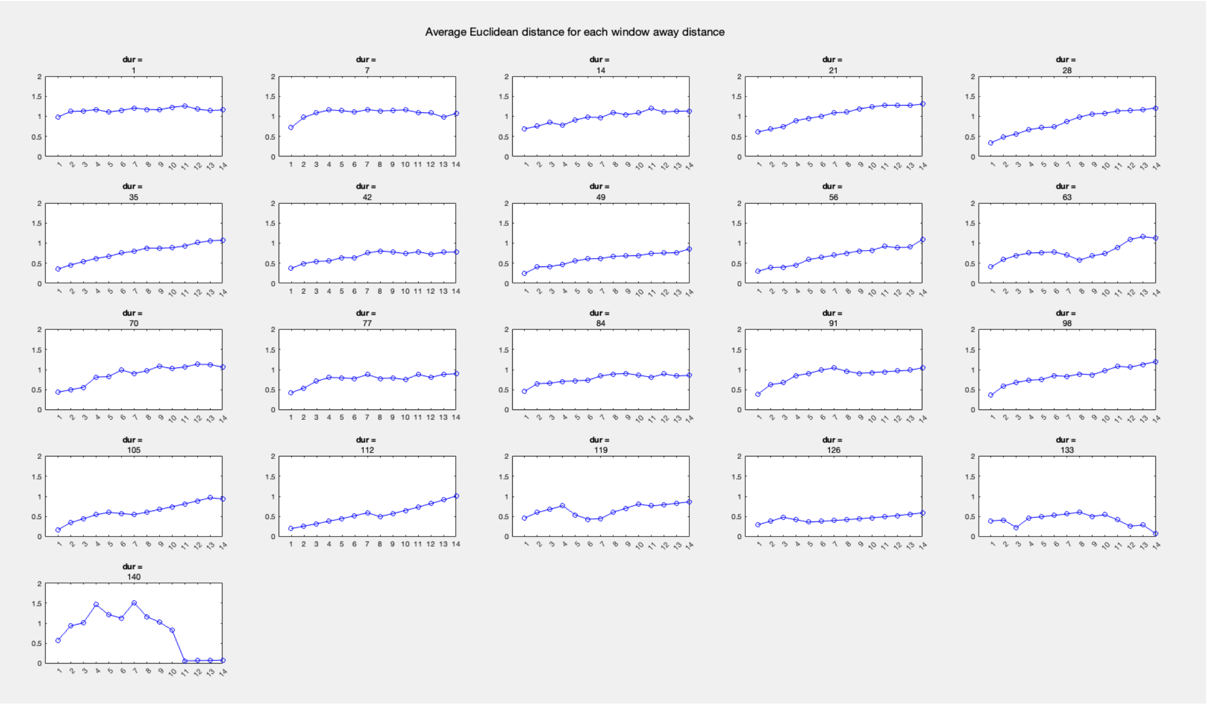
Selecting the Optimal Window-away Distance

To evaluate the physical similarity of the normalized proximity matrices generated by the different averaging duration, the Euclidean distance between each normalized proximity matrix's first set of eigenvectors (the eigenvectors associated with the largest eigenvalue) is calculated. Similar to the above procedure, the average Euclidean distance of the matrices (or, the average of the L2 norm) generated by different window-away distances was calculated for each averaging duration. Specifically, the average Euclidean distance E_n for each window-away distance n given x within-bin window pairs' proximity matrices' first eigenvectors sets, where each pair of first eigenvector set is denoted as (x_i , x'_i), is calculated as the following,

$$E_n = \frac{\sum_{i=0}^{x} ||x_i - x'_i||_2}{x} .$$

Fig. 4. presents the average Euclidean distance E_n of windows in each window-away distance at each averaging duration. Consistent with the average of the modified rand index of the spectral clustering result, the average Euclidean distance of windows separated by each window-away distance at each averaging-over duration also shows the optimal window-away distance is 1 window away as the value of E_1 is lowest across durations. In other words, the Euclidean distance, or the physical similarity, between the proximity matrices derived from the time windows that are separated by one window are overall similar to each other across all averaging-over durations (or the window size). With this optimal window, we are able to compare the durations at their most stable comparison rates of the windows.

Figure 4. Average Euclidean distance *E_n* for the first eigenvectors set of the proximity matrices derived from the windows that are distanced away by each window-away distance *n* at each averaging duration *dur.* The average Euclidean distance E_n for the first eigenvectors of the proximity matrices at each window-away distance is calculated by taking the average of the Euclidean distance E_n between each two proximity matrices that are separated by n windows. Although the distribution of E_n varies slightly for duration = 135 days and 140 days, the lowest Euclidean distance E_n occurs around window-away distance n = 1.



Selecting the Optimal Averaging Duration

With the average Euclidean distance E_n of the normalized proximity matrices' first eigenvectors generated from windows that are separated by different window-away distances n at different duration dur, we compare average E_n for each window-away distance across all averaging-over durations to evaluate the physical similarity between the proximity matrices generated by the different averaging duration. By this comparison, we aim to find the averaging-over duration that has the lowest average Euclidean distance across all different window-away distances which indicates the proximity matrices of the time windows that are associated with this averaging-over duration are relatively stable and therefore, optimal for representing the aggregated proximity matrix of the whole data collection period. In other words, the identified optimal averaging-over duration provides an ideal duration that is capable to produce the social structure that is relatively accurate to represent the overall social structure of the macaque during the data collection period without using extensive data.

We expect a relatively lower value of the average Euclidean distance E_n across all window-away distances n for an optimal averaging-over duration as it will indicate the relative physical stability of the normalized proximity matrices. We will especially focus on examining the fluctuation of the average Euclidean distance across the duration when the window-away distance = 1 window as the proximity matrices from the windows that are separated by this distance generally have a relatively small difference across all durations. We also expect the Euclidean distance E_n of the optimal averaging-over duration to be at an elbow point where the fluctuation, or decrease, of E_n begins to decline.

Comparing the Results With the Actual Matrilineal Genetic Tree

To verify the accuracy of the clustering result based on the optimal averaging-over duration, a similarity matrix based on the genetic similarity of each dyad of macaque was built. The genetic similarity was calculated based on the matriline connection of the 26 macaques. The genetic similarity score s between individual x and itself is measured as 1 $(i.e., \ s(x,\ x)\ =\ 1)$. Each n connection away individual y will have the genetic similarity score *s* with *x* being$s(x,\ y)\ =\ 0.5^{n}$. Therefore, the genetic similarity score of one's parents and siblings is measured as 0.5 and the similarities to one's cousins are measured as 0.25. Appendix A shows the calculated similarity matrix based on genetic similarity. A spectral clustering with $k = 6$ was performed based on the genetic similarity matrix as shown in Fig. 5. The following is the resulting cluster:

Cluster 1: {Nl16, 151MX, Mw16, Pw13}

Cluster 2: {Cn17, Ki18, Sr11, Nm17, Zc18, Iv11, Dt13, Do12, Fs16, Cc11, Gh10}

Cluster 3: {Aq15, Gm17, Ih11}

Cluster 4: {Id18, Nk15, Mm17, Mw13}

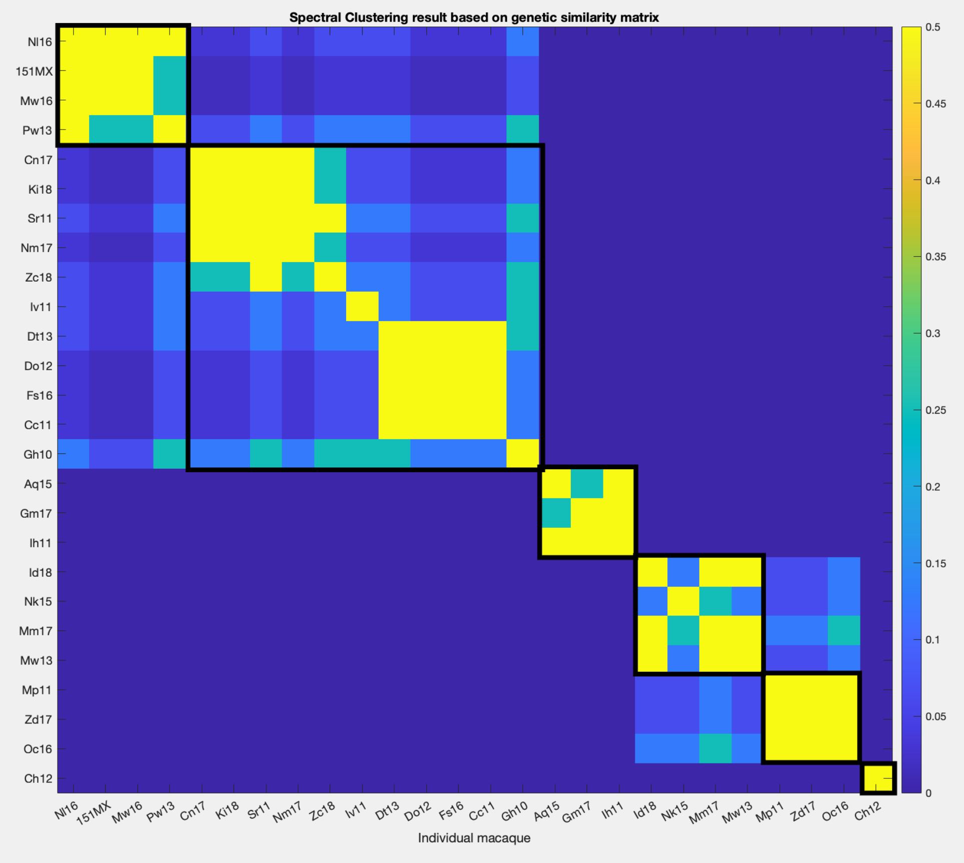
Cluster 5: {Mp11, Zd17, Oc16}

Cluster 6: {Ch12}

The result of the spectral clustering based on the genetic similarity matrix was compared with the original and shuffled spectral clustering result generated at k = 6 across all windows in all averaging-over durations by computing the modified rand index between them. By computing the modified rand index, we are able to identify the averaging-over duration that has the most similar clustering result with the clustering result of the genetic matrix on average. In other words, we will obtain the averaging-over duration that can generate the social structures based on the proximity matrix of each of its windows that are most similar to macaques' actual matrilineal family structure on average.

Since the 3rd window was recognized as the optimal window representing all windows as aforementioned, we also compared the 2-norm of the pairwise Euclidean distance between the genetic similarity matrix and the normalized proximity matrices generated at the 3rd window across each averaging-over duration to investigate the similarity between the two. The shuffled normalized proximity matrices at the 3rd window of each duration were also compared across different durations with the genetic similarity matrix by finding the 2-norm of the pairwise Euclidean distance between them. By doing so, we can verify the previous comparison result. However, since the genetic similarity matrix is only based on the matriline family tree and our data include male macaques and the matrix only contains an approximate discrete value of the pairwise genetic similarity, the comparison can produce an inconsistent result.

Figure 5. Spectral clustering result based on the genetic similarity matrix. Cluster is generated by spectral clustering on the genetic similarity at the number of clusters *k = 6.*



Results

Background

The adaptation of the tracking device enables primate studies to obtain more comprehensive and longitudinal behavior data of the animals. Nevertheless, such data often comes with enormous computational challenges. In the previous relevant studies, it was discovered that data obtained at a very short timescale is comparable with the actual human observational data (Gelardi et al., 2020). In addition to the computational challenges, few studies focused on using longitudinal tracking data to generate and infer the social network of the rhesus macaque. Therefore, this study is aimed to evaluate the accuracy vs. temporal resolution for generating the social network of Rhesus Macaque with 154 days of data collected from 2021 March 4th to 2021 August 4th. It specifically investigates the optimal averaging-over duration among *duration = {1 day, 7 days, 14 days, 21 days …, 140 days}* for generating a stable network that can relatively represent the network generated from the long-term timescale.

Results of Spectral Clustering Evaluated by Modified Rand Index

By calculating the average modified rand index $W_{n\ at\ duration\ =\ dur}$ of each window-away distance *n* $\in \{1, 2, 3, \ldots, 14\}$ for each averaging-over duration *dur* $\in \{1\ day, 7\ days, \ldots, 140\ days\}$, we obtain Fig. 6. The x-axis in Fig. 6. shows the averaging-over duration and the y-axis shows the average modified rand indexes of each averaging-over duration. Lines labeled by “o” are the average modified rand index of non-shuffled clustering results across each averaging-over duration whereas the lines labeled by “x” are the average modified rand index of shuffled clustering results. Different window-away distances are distinguished by color labeled in legend.

The average modified rand index of the non-shuffled clustering results shows an increasing pattern from averaging-over duration = 1 day to 14 days. The increasing pattern becomes stable after averaging-over duration = 14 days. For window-away distance = 1 window (colored by the blue line labeled as 1), the averaging modified rand index fluctuates within ± 0.05 after reaching $W_n = 0.95$ at averaging-over duration = 14 days. Before this elbow point, the averaging modified rand index increased from 0.73 to 0.95. Similar patterns were observed for all other window-away distances. However, for window-away distances 7 to 14 windows, there is over 0.05 increase in the average modified rand index after averaging-over duration = 98 days. The increase slows and decreases slightly within 0.05 after reaching averaging modified rand index ≥ 0.95 at averaging-over duration = 112 days. The average modified rand indexes of the shuffled clustering results are generally below the result of the non-shuffled clustering results. None of the fluctuation in the shuffled result across all different averaging-over duration in each window-away distance is above ±0.05.

The similarity of the clustering results of all windows increased greatly from averaging-over duration = 1 day (modified rand index = 0.73) to averaging-over duration = 7 days (MRI = 0.90) to averaging-over duration = 14 days (MRI = 0.95). The increase of stableness drops as the averaging-over duration reaches 14 days with the MRIs of the subsequent duration on the average increase within 0.03.

Figure 6. Average modified rand indexes of windows that fall into the bin in each averaging-over duration for each window-away distance. The average MRI of each duration at each window-away distance is calculated by taking the average of the modified rand indexes of the clustering results of the windows that are separated by the window-away distance. The elbow point of the distribution occurs around the averaging-over duration = 14 days.

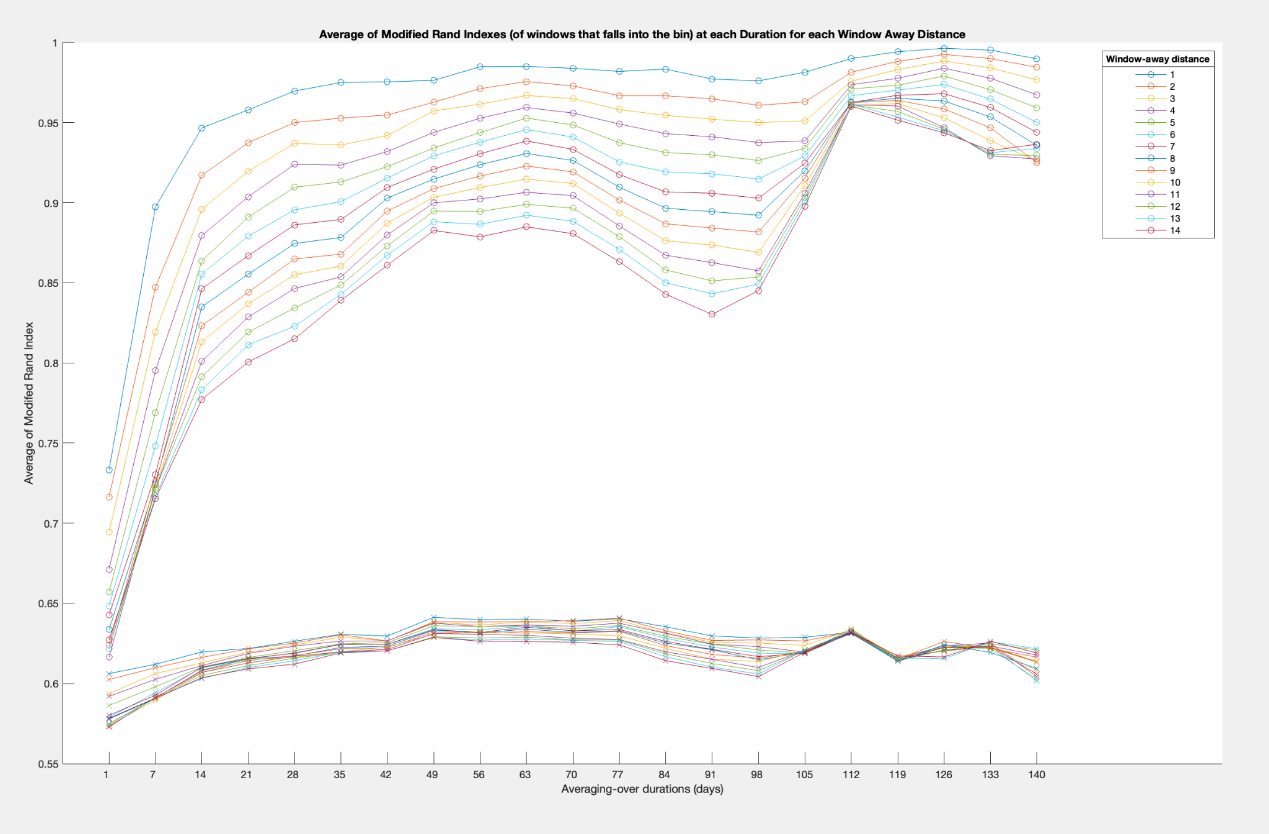


Fig. 7. presents the six clusters resulting from the spectral clustering at k = 6, window = 3, and averaging-over duration = 14 days. The result presents the following clusters,

Cluster 1: {Cn17, Ki18, Nm17, Sr11}

Cluster 2: {151MX, Mw16, Pw13}

Cluster 3: {Nl16}

Cluster 4: {Cc11, Ch12, Do12, Dt13, Fs16, Gh10, Id18, Mm17, Mp11, Mw13, Nk15, Oc16, Zd17}

Cluster 5: {Gm17}

Cluster 6: {Aq15, Ih11, Iv11, Zc18} .

Fig. 8. presents the average modified rand index between the clustering result (original and shuffled) based on the proximity matrices generated with different averaging-over durations across all its windows and the clustering result based on the genetic similarity matrix. The MRI peaked around averaging-over duration = 14 days, which is consistent with the average MRI of the clustering result of windows. The highest modified rand index between the original spectral clustering results based on the proximity matrix and the genetic similarity matrix occurs at duration =14 days with MRI = 0.62. The error bars are calculated by the standard deviation of the rand indexes of windows within individuals' duration compared to the clustering result of the genetic similarity matrix.

Fig. 9. shows the 2-norm of the pairwise Euclidean distance between the genetic similarity matrix and the proximity matrix (at the 3rd window) generated across the averaging-over duration. All shuffled proximity matrices have a higher value of the 2-norm. The lowest 2-norm occurs when averaging-over duration = 1 day. However, the overall differences between the Euclidean distances at each averaging-over duration are within ±2.

Figure 7. Spectral clustering results at k = 6, window = 3rd, averaging-over duration = 14 days. The spectral clustering is based on the aggregated proximity matrix generated by the data from 2021/03/06 to 2021/03/19.

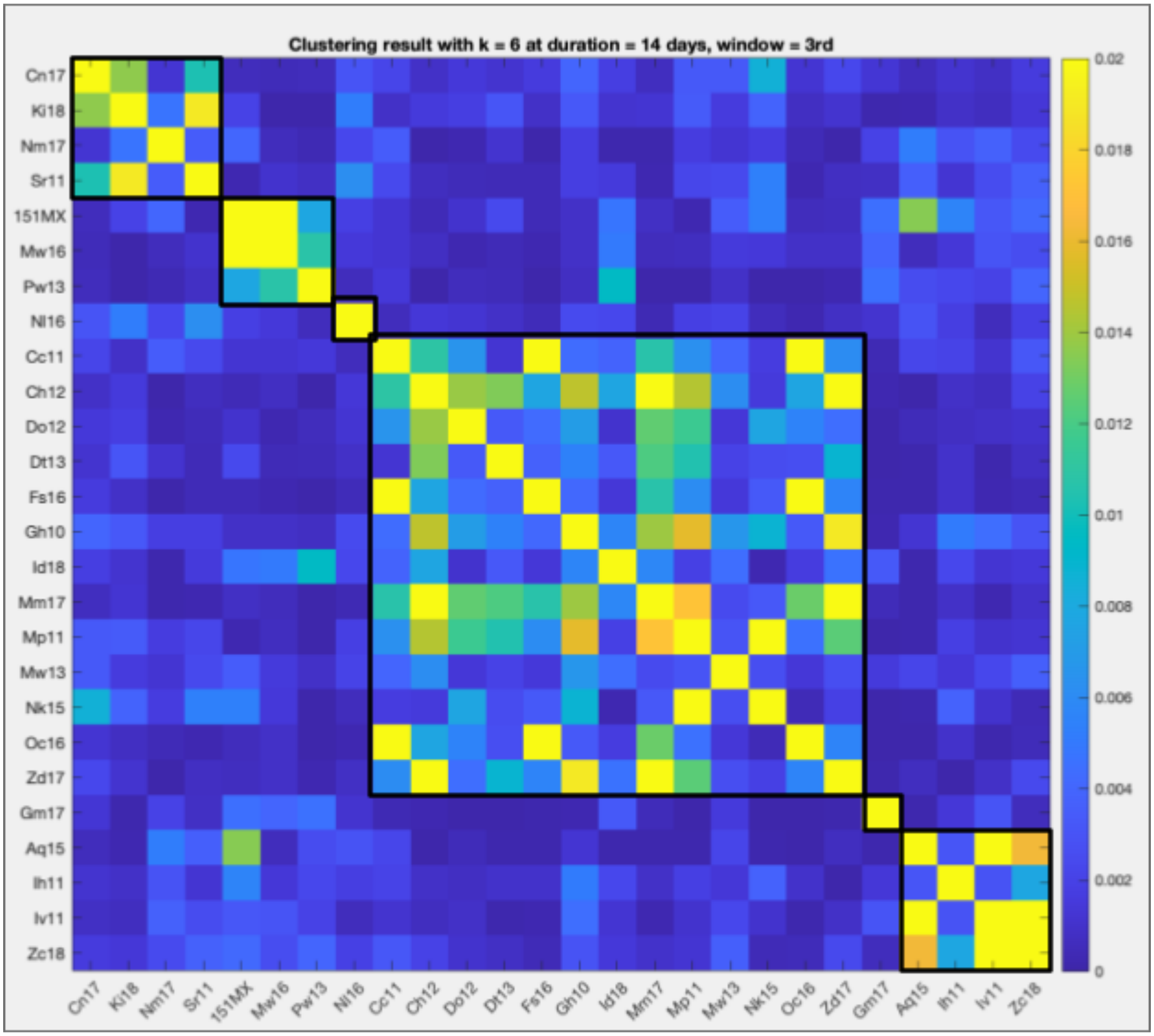


Figure 8. Average of modified rand index between the spectral clustering result based on the genetic similarity matrix and the proximity matrices that are generated by all windows at different averaging-over durations. The average modified rand index at each averaging-over duration is calculated by taking the average of the modified rand indexes between the spectral clustering result of the genetic similarity matrix (k = 6) and the clustering results of the normalized proximity matrices generated by each window at the averaging-over duration. The shuffled clustering result indicates the normalized proximity matrix-based spectral clustering result was shuffled and then compared with the clustering result of the genetic similarity matrix. The error bar is calculated by the standard deviation of the modified rand indexes at each averaging-over duration. The averaging-over duration that has the highest value of the average modified rand index with the non-shuffled clustering result is 14 days.

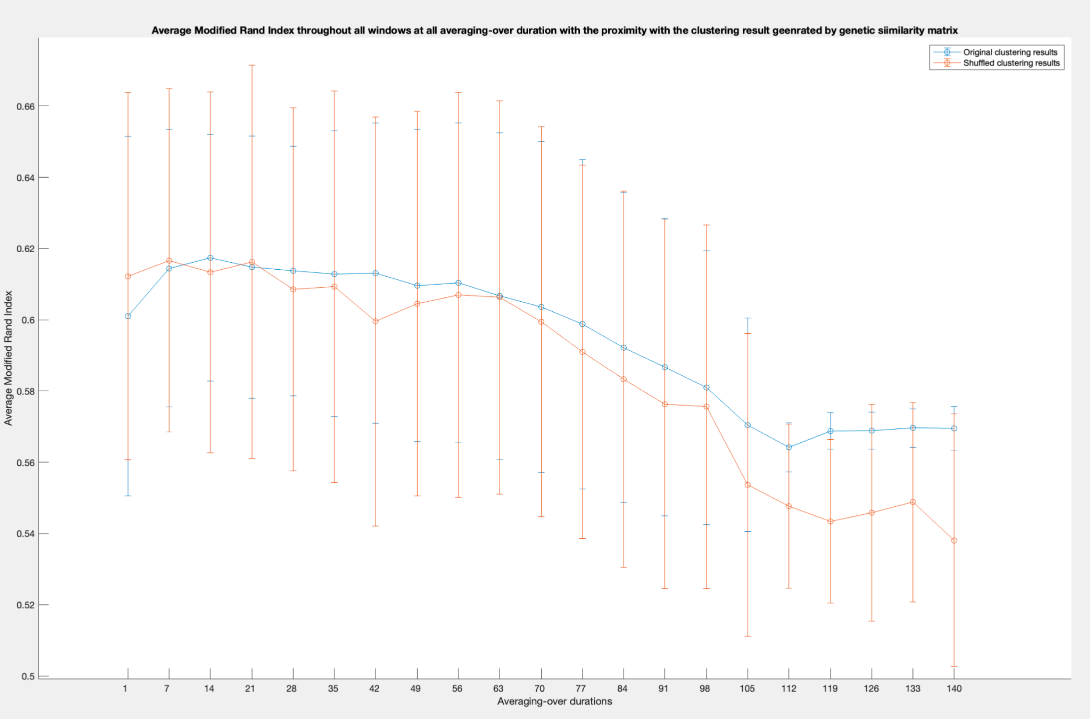
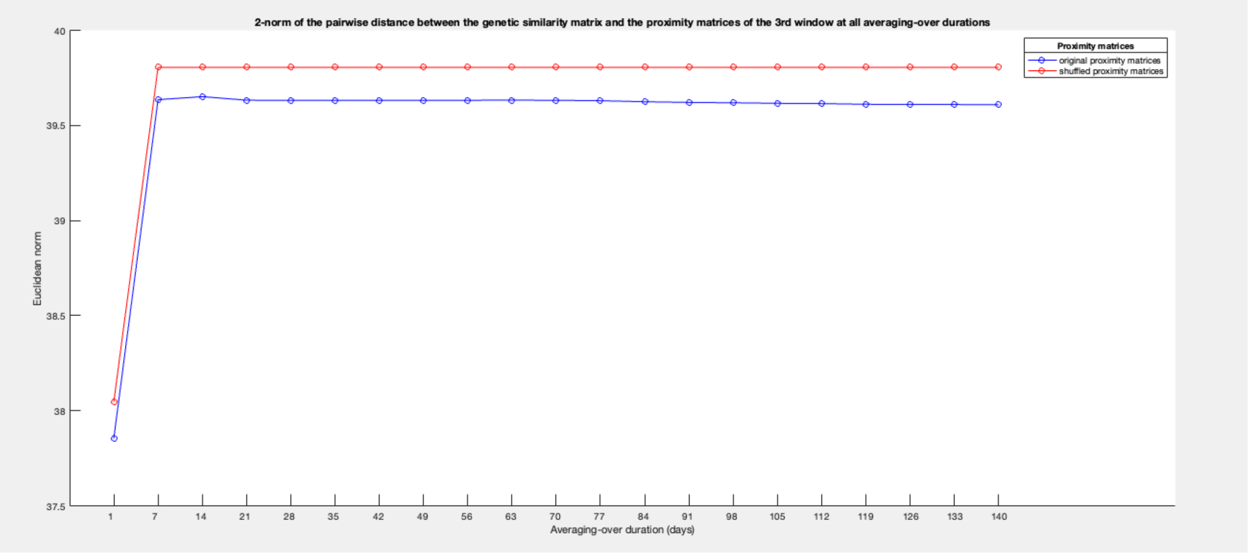


Figure 9. 2-Norm of the pairwise Euclidean distance between the genetic similarity matrix and the proximity matrices (at the 3rd window) generated across averaging-over durations. The 2-norm of the pairwise Euclidean distance at each averaging-over duration was computed by finding the square root of the largest eigenvalue of the pairwise distance matrix's inner product matrix. The pairwise distance matrix is computed by finding the Euclidean distance between the normalized proximity matrix of the 3rd time window at each duration and the genetic similarity matrix. The lowest 2-norm of the pairwise distance occurs at duration = 1 day. The magnitude, or the 2-norm, of the distance between the two matrices, stays consistent after duration = 1 day.



Results of Euclidean Norm of the First Eigenvectors

By computing the average Euclidean distance E_n of the first eigenvector sets associated with normalized proximity matrices that are generated from the window pairs that fall into each window-away distance $n \in \{1, 2, 3, \ldots, 14\}$ for each averaging-over duration $dur \in \{1\ day, 7\ days, \ldots, 140\ days\}$, we obtain Fig. 10. The x-axis in the figure denotes the averaging-over durations and the y-axis in the figure denotes the average Euclidean distance $E_{n.}$ Different window-away distances are distinguished by color labeled in legend. The average Euclidean distance shows a decreasing pattern from averaging-over duration = 1 day to 49 days across all different window-away distances. For window-away distance = 1 window, the lowest Euclidean distance E_1 occurs at averaging-over duration = 105 ($E_n = 0.17$). The elbow point of the trend occurs around averaging-over duration = 28 days, where the changes after the point remain within the interval of ±0.4. For window-away distance other than one window, the average Euclidean distance E_n fluctuates across the averaging-over durations within the interval of ±0.8. Throughout all window-away distances, the average Euclidean distance E_n reaches the relatively lowest value around duration = 49 days.

Social Network Generated Based on the Selected Data

Fig. 11. and Fig. 12. present the social network generated by the normalized proximity matrices with the averaging-over duration = 14 days and window = 3rd. The edge width represents the proximity between the dyads.

Figure 10. Average L2 norm (Euclidean distance) of the first eigenvectors of each averaging-over duration at each window-away distance. The average L2 norm of the first eigenvectors at each averaging-over duration and window-away distance is calculated by taking the average of the L2 norm of the first eigenvectors of the proximity matrix from the windows separated by the window-away distance at the duration. The value of the L2 norm is the lowest for window-away distance = 1 window around duration = 28 days.

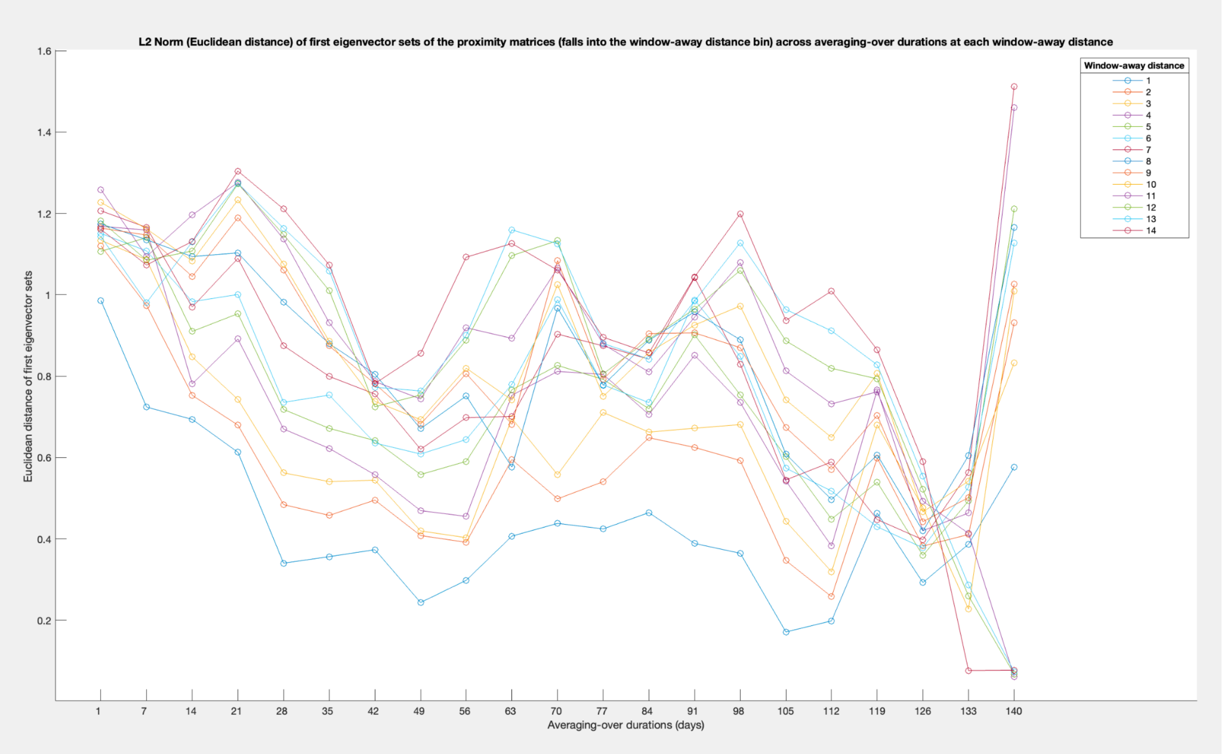


Figure 11. The social network of 26 rhesus macaques based on their normalized aggregated proximity score from 2021/03/06 to 2021/03/19

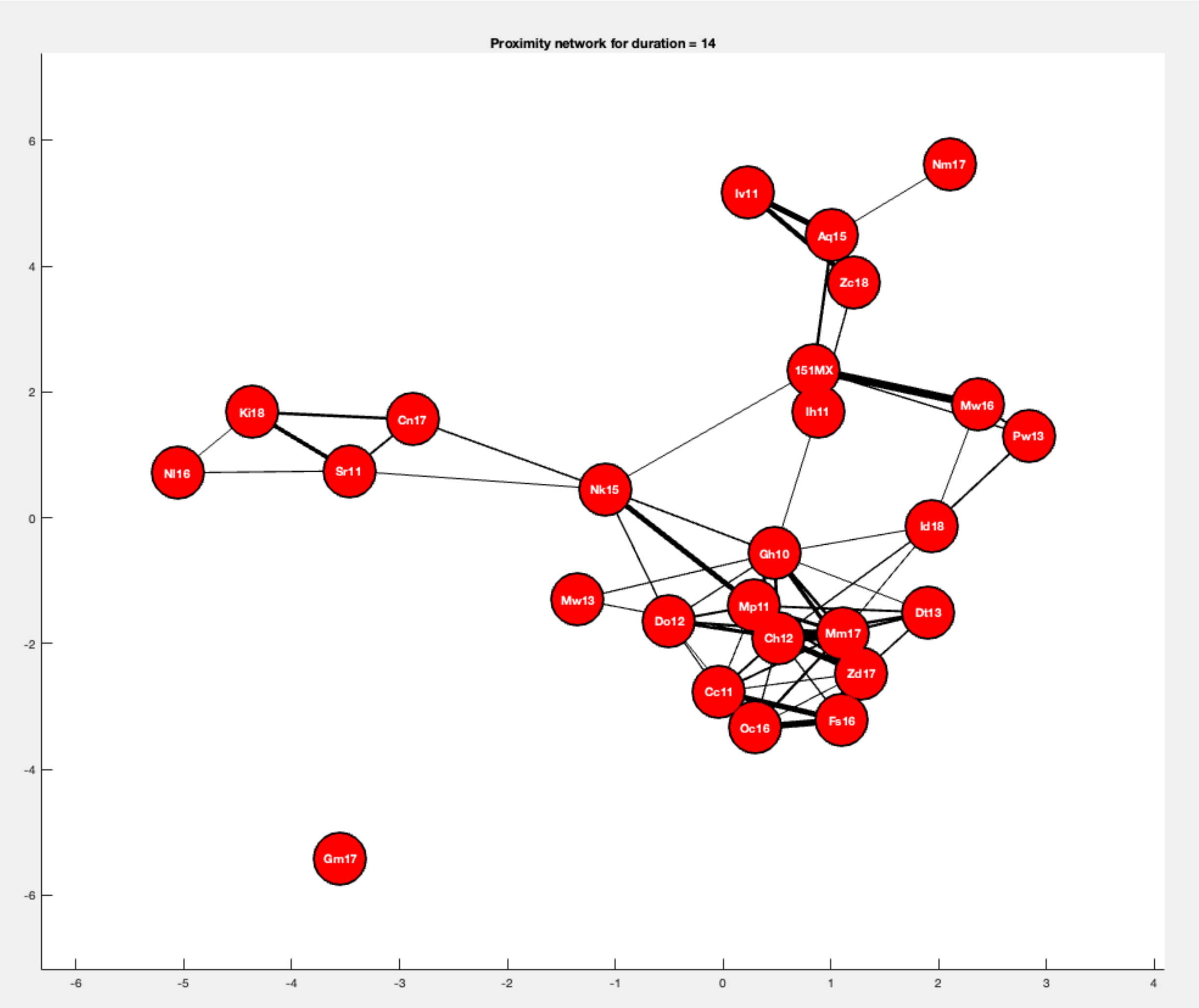
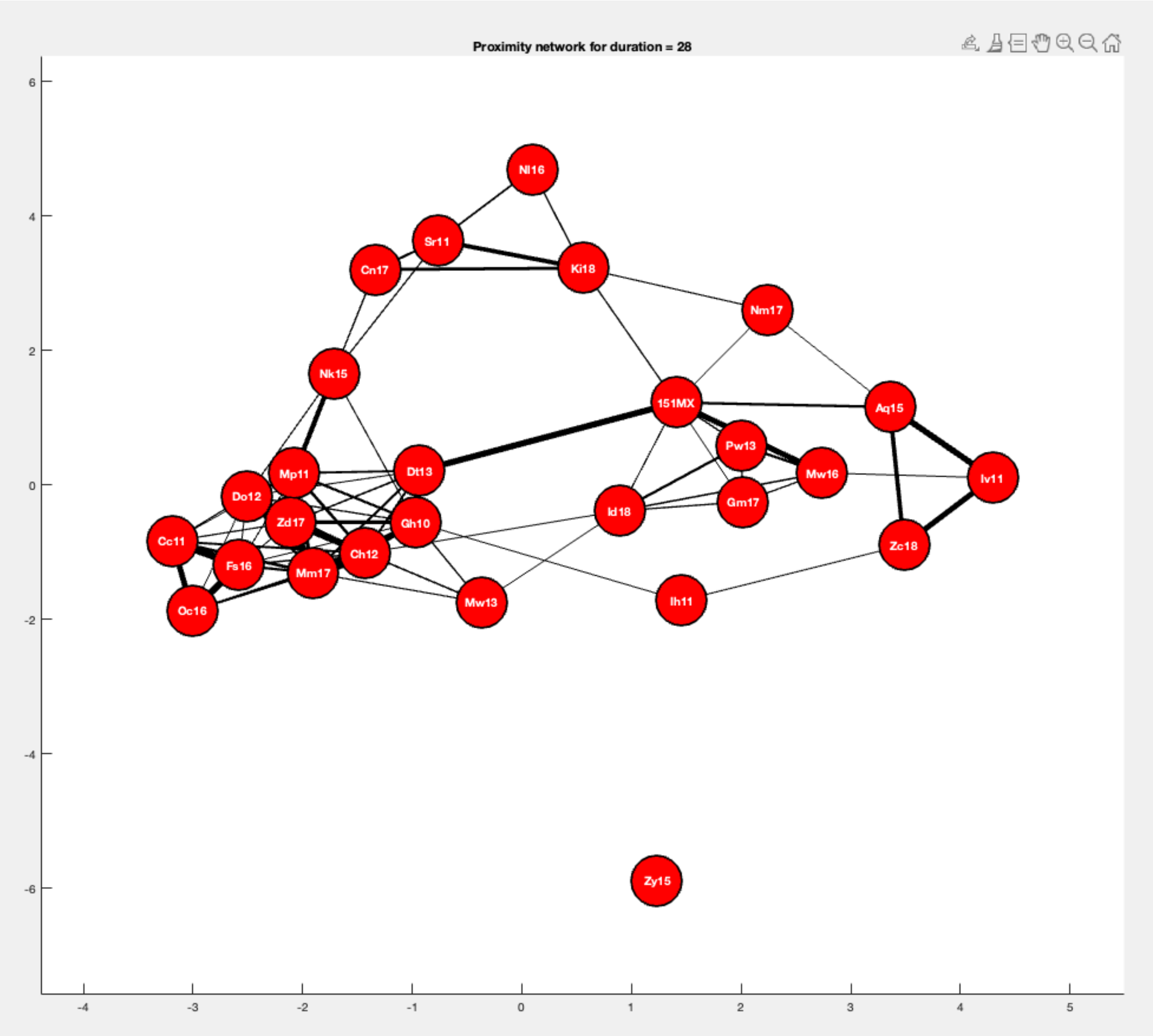


Figure 12. The social network of 26 rhesus macaques based on their normalized aggregated proximity score from 2021/03/06 to 2021/04/02



Discussion

The macaque species share important social cognition and behavior characteristics with humans. As tightly connected as their social structure with their social behavior, studying the social structure of macaques can reveal significant insights into human psychology and behavior, such as social learning capabilities (Fiore et al., 2020). In addition to their significance in social behavior and psychology studies, they are also critical model organisms for modeling infectious disease transmission of their great similarity to human physiology, neurobiology, and susceptibility. Therefore, it becomes critical to acquire accurate knowledge of their social structure. With the advancement of technology, animal behavior studies are able to collect continuous high-resolution positional data with tracking devices that can overcome the time limitations of traditional observational studies. However, such data collection methods often come with huge computational challenges. Since very few studies in the past have used longitudinal tracking-device data to infer the social network of rhesus macaque and little focused on finding the solution for resolving the computational challenges in using such data in inferring macaque's social structure, this study focuses on examining the social network generated with different averaging-over durations. It specifically focuses on analyzing the networks generated with different short-term time frames within the half-year data collection period (from 2021/03/04 to 2021/08/04) and determining the optimal averaging-over duration for generating a stable and accurate social network of the sampled rhesus macaques. In summary, the result suggests sampling 14 days or 28 days are the potential optimal averaging-over durations for generating a stable network that can represent the whole period from 2021/03/04 to 2021/08/04.

With the normalized proximity matrices generated from different averaging-over durations of 1 day, 7 days, …, 140 days, we obtained different spectral clustering results at the number of clusters k= 6. By computing the average modified rand indexes of the clustering results of windows that are separated by each window-away distance across all averaging-over duration, we found the elbow point of increasing trend occurs at the 14-day duration over all window-away distances from 1 to 14 windows as shown in Fig. 6., indicating using 14 days data throughout different time windows can generate a relatively stable proximity matrix or the social network of macaques, that can relatively well-represents their social structure in the whole short-term period of 154 days with the minimum amount of data. For instance, the spectral clustering's results of the normalized aggregated proximity matrix generated by averaging over 14-days duration stay more consistent throughout the time windows {(2021/03/04 to 2021/03/17), (2021/03/05 to 2021/03/18), …} than the clustering results generated by all the time windows that are averaging over the 1-day duration, which include the window {2021/03/04, 2021/03/05, …}. In other words, sampling randomly 14 consecutive days of data can produce a social structure that is relatively well-representing the social structure of the entire 154-day data collection period since the normalized aggregated proximity matrix of 14-day averaging-over duration produce the spectral clustering results that have the highest similarity, or stability, with each 14-days time window.

As shown in Fig. 6., for window-away distances = 6 windows to 14 windows, there is another elbow point around averaging-over duration = 112 days where the average rand index increase ≥ 0.05 in the previous three consecutive durations. This might be due to the effect of the sample size decrease for these window-away distances as the slope of change of the average modified rand index becomes steeper as the window-away distance becomes larger around averaging-over duration = 112 days. The overall lower average modified rand index for larger window-away distance across all durations also demonstrates this point. Compared to the original clustering results, the shuffled clustering results consistently produced a lower value of modified rand indexes with each window-away distance throughout the averaging-over durations. Therefore, the distribution of the modified rand index of the original clustering result is not generated by chance. The stableness of the social network, or the clustering results, measured by the proportion of similar clustering between the time windows of the duration, increased greatly from averaging-over duration = 1 day (MRI = 0.73) to averaging-over duration = 7 days (MRI = 0.90) to averaging-over duration = 14 days (MRI = 0.95). The increase rate of stableness drops as the averaging-over duration reaches 14 days with the proportion of similar clustering of the subsequent duration on the average increase within 0.03. Overall, comparing the average rand index of the spectral clustering at different window-away distances across all different averaging-over duration shows averaging over 14 days of data is the optimal point for generating the social network of rhesus macaques sampled in this study.

Apart from the average modified Rand indexes between proximity matrices, the study also examined the average modified Rand indexes between the spectral clustering results of the genetic similarity matrix and proximity matrices generated from all windows across the averaging-over durations (as shown in Fig. 8.). The results indicate that the duration of 14 days has the highest MRI, suggesting that the normalized proximity matrices generated from 14 days of data can effectively represent the genetic similarity matrix. This finding confirms that the social structure inferred from 14 days of data can well-represent the social structure during the 154-day data collection period on average. The study found that the similarity between the proximity matrix-based spectral clustering result and the clustering result of the genetic similarity matrix increased from duration = 1 day to 14 days (MRI = 0.6 to 0.61) and then decreased from duration = 14 days to 140 days (MRI = 0.56) with a change of 0.1. The average modified rand index between the clustering result of the genetic similarity matrix and the shuffled clustering results from the normalized proximity matrices produced results that are overall lower than the values sampled from the original proximity matrices which shows the finding is not necessarily due to randomness. However, it should be noticed that the physical values of changes of the average MRI across all durations are relatively small (± 0.1) and the value is consistently located within the interval of 0.52 to 0.62 which might potentially be due to the overgeneralization of the entries in the genetic similarity matrix. For example, macaque 151MX was not in the matriline family tree and therefore, he will have zeros for all the connections to female macaques in the genetic similarity matrix. Nevertheless, 151MX has a strong bond, or high proximity score, with Mw16 throughout the data sampling period. Therefore, the inaccurate or overgeneralized genetic similarity score between dyads may lead to a lower value of the average modified index.

Similarly to the spectral clustering results, the magnitude (2-norm) of the pairwise Euclidean distance between the genetic similarity matrix and the proximity matrices (at the 3rd window, selected by the analysis described in the method section) across averaging-over durations shows very small changes throughout the durations as shown in Fig. 9. Only results from one-day duration to seven-day duration have come through an increase of 1.5. After duration = 7 days, the magnitude of the pairwise Euclidean distances between the genetic similarity matrix and the proximity matrices stay consistently around 39.5. The shuffled proximity matrices also consistently have a higher value of Euclidean distance compared to the original proximity matrices, indicating the distribution was not due to randomness. Therefore, both the average modified rand index between the spectral clustering result of the genetic similarity matrix and the proximity matrices generated by different averaging-over duration and the Euclidean distance between the genetic similarity matrices and proximity matrices (Fig. 6., Fig. 8., Fig. 9.) demonstrates two points:

1. The proximity matrices generated from different averaging-over durations of 1 day, 2 days,...140 days all can approximately represent 50% to 60% of the overall matriline structure. In other words, the proximity between macaques follows about 50% to 60% of the matriline genetic closeness pattern.
2. The proximity between macaques, or their social structure, stays relatively stable throughout different averaging-over durations in terms of their deviation from their actual genetic similarity tree.

The above two points may ultimately imply the relatively stable nature of rhesus macaque's social structure in terms of how they follow the matriline genetic closeness for the affiliative behavior of staying in close proximity with others. Indeed, female rhesus macaques show a stable preference for associating with kin (Beisner et al., 2010). Nevertheless, the closed matrilines defined by genetic connection can be broken down by events such as mothers of individuals who are closely related succumbing to predation, old age, or illness (Beisner et al., 2010). Such fragmentation of matrilines might as well explain the relatively medium similarity between the proximity matrices and genetic similarity matrices across all different durations. Regardless of the stable patterns overall, the averaging-over duration = 14 days still shows a relatively higher index for the metrics and therefore, indicating it can effectively represent the social structure of the macaques during the whole data collection period.

Besides the comparison with the matriline genetic similarity matrix, the study also compared the average Euclidean distance of the proximity matrices generated by each averaging-over duration at each window-away distance as shown in Fig. 10. Different from the distribution from the previous analysis, the Euclidean distance of the proximity matrices' first eigenvectors of the windows falling into the window-away distance at each duration has the elbow point of decreasing at duration around 28 days across window-away distance 1 to 6 windows (Fig. 10.), which indicates 28 days of data can acquire us a relatively stable proximity matrix. For window-away distance = 7 windows to 14 windows, the elbow point shifts from 28 to 42 days averaging over durations. Since our previous analysis on window-away distance suggests window-away = 1 window returns the relatively most stable output, we will recognize 28 days duration as the relatively optimal point for the averaging-over duration suggested by the Euclidean distance.

The different optimal averaging-over duration suggested by the two analyses, Euclidean distance between eigenvectors of proximity matrices generated by different durations, and the spectral clustering result modified rand index across the different durations, might be due to the underlying methodological and algorithmic differences of the comparison of physical matrices and the clustering results. Since the Euclidean distance to the genetic matrix shows very little difference between the two durations, it is necessary to have observational data to compare and validate the two networks generated by the two different averaging durations. Nevertheless, there were no currently validated social networks constructed based on the observational data for the 26 macaques. Therefore, the study presents the generated networks based on the two potential optimal averaging-over durations in Figures 11 and 12 and their associated descriptive network measures in Appendix B for future analysis.

In conclusion, through conducting spectral clustering and analyzing the modified rand index between windows falls into the same window-away distance bin at each averaging-over duration (as shown in Fig. 6.) we were able to find the optimal averaging-over duration is 14 days. This result was also confirmed by the analysis of the average modified rand index between the clustering results from genetic similarities matrices and proximity matrices generated across durations (as shown in Fig. 8.). A nearby duration of 28 days was also suggested to be optimal according to the Euclidean distance of the first eigenvector sets of the proximity matrices from the windows falls into each window-away distance bin across the averaging-over durations (as shown in Fig. 10.). Therefore, this study found 14 days and 28 days of data can produce a social network that can relatively represent the social network during the whole data collection period of 154 days compared to other durations. The social structure produced by 14 days of data on average has a higher similarity with the actual matrilineal genetic similarity structure.

Limitations and Future Directions

Due to the time limitation, data were only sampled for a five-month period and analyzed with 1-day, 7-day, …, and 140-day averaging-over durations. If time allows, sampling data for all three years data sampling period 2018/08 to 2022/01 will reveal more insights into the changes in the network over time since it permits potential time for detecting the actual social structure change among the macaques. The lack of the observational data constructed network also limited the capability of this study to verify its finding. Also, the genetic matrix only contains matriline females, and therefore, the comparison result between it and the proximity matrices in the study might not be accurate. Therefore, the future research direction can be analyzing the accuracy vs. temporal resolution of social networks generated from longitudinal data by comparing the result with actual social networks built from the observational data to discover the true optimal sampling rates. It will also be interesting to evaluate and compare the dynamic network of different non-human primate species generated from longitudinal data to investigate how network changes can vary across species. From the network analysis perspective, changing the network to directed edges, and adding the parameter of gender and age might all provide significant insights into understanding the network changes over time, the social structure of the analyzed species, and the hierarchy of the primates.

References

Brent, L. J. N., Heilbronner, S. R., Horvath, J. E., Gonzalez-Martinez, J., Ruiz-Lambides, A., Robinson, A. G., Skene, J. H. P., & Platt, M. L. (2013). Genetic origins of social networks in rhesus macaques. *Scientific Reports*, *3*(1). https://doi.org/10.1038/srep01042

Chang, E.-C., Huang, S.-C., & Wu, H.-H. (2009). Using K-means method and spectral clustering technique in an outfitter's value analysis. *Quality &amp; Quantity*, *44*(4), 807–815. https://doi.org/10.1007/s11135-009-9240-0

Drayton, L. A., & Santos, L. R. (2014). A decade of theory of mind research on cayo santiago: Insights into rhesus macaque social cognition. *American Journal of Primatology*, *78*(1), 106–116. https://doi.org/10.1002/ajp.22362

Ehardt, C. L., & Bernstein, I. S. (1986). Matrilineal overthrows in rhesus monkey groups. *International Journal of Primatology*, *7*(2), 157–181. https://doi.org/10.1007/bf02692316

Fiore, A. M., Cronin, K. A., Ross, S. R., & Hopper, L. M. (2020). Food cleaning by Japanese macaques: Innate, innovative or cultural? *Folia Primatologica*, *91*(4), 433–444. https://doi.org/10.1159/000506127

Gardner, M. B., & Luciw, P. A. (2008). Macaque models of human infectious disease. *ILAR Journal*, *49*(2), 220–255. https://doi.org/10.1093/ilar.49.2.220

Gelardi, V., Godard, J., Paleressompoulle, D., Claidière, N., & Barrat, A. (2020). *Measuring social networks in primates: Wearable sensors vs. direct observations*. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/2020.01.17.910695

Gibbs, R. A., Worley, K. C., Kehrer‐Sawatzki, H., & Cooper, D. N. (2008). The Sequencing of the Rhesus Macaque Genome and its Comparison with the Genome Sequences of Human and Chimpanzee. *eLS*. https://doi.org/10.1002/9780470015902.a0020744

Hrolenok, B., Balch, T., Byrd, D., Roberts, R., Kim, C., Rehg, J. M., Gilliland, S., & Wallen, K. (2018, December 4). Use of position tracking to infer social structure in rhesus macaques.

Proceedings of the Fifth International Conference on Animal-Computer Interaction. http://dx.doi.org/10.1145/3295598.3295613

Kalin, N. H., & Sheltona, S. E. (2003). Nonhuman primate models to study anxiety, emotion regulation, and psychopathology. *Annals of the New York Academy of Sciences*, *1008*(1), 189–200. https://doi.org/10.1196/annals.1301.021

Keeling, M. J., & Eames, K. T. D. (2005). Networks and epidemic models. *Journal of The Royal Society Interface*, *2*(4), 295–307. https://doi.org/10.1098/rsif.2005.0051

Lewis, A. D., & Prongay, K. (2015). Basic Physiology of Macaca mulatta. In *The Nonhuman Primate in Nonclinical Drug Development and Safety Assessment* (pp. 87–113). Elsevier. http://dx.doi.org/10.1016/b978-0-12-417144-2.00006-8

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/tit.1982.1056489

Maddali, H. T. (2014). Inferring Social Structure and Dominance Relationships Between Rhesus macaques using RFID Tracking Data. *Georgia Institute of Technology*, *x*(x).

Maestripieri, D., & Hoffman, C. L. (2011). Behavior and social dynamics of rhesus macaques on Cayo Santiago. In *Bones, Genetics, and Behavior of Rhesus Macaques* (pp. 247–262). Springer New York. http://dx.doi.org/10.1007/978-1-4614-1046-1_12

McCowan, B., Beisner, B. A., Capitanio, J. P., Jackson, M. E., Cameron, A. N., Seil, S., Atwill, E. R., & Fushing, H. (2011). Network stability is a balancing act of personality, power, and conflict dynamics in rhesus macaque societies. *PLoS ONE*, *6*(8), e22350. https://doi.org/10.1371/journal.pone.0022350

Monfardini, E., Reynaud, A. J., Prado, J., & Meunier, M. (2017). Social modulation of cognition: Lessons from rhesus macaques relevant to education. *Neuroscience &amp; Biobehavioral Reviews*, *82*, 45–57. https://doi.org/10.1016/j.neubiorev.2016.12.002

Munster, V. J., Feldmann, F., Williamson, B. N., van Doremalen, N., Pérez-Pérez, L., Schulz, J., Meade-White, K., Okumura, A., Callison, J., Brumbaugh, B., Avanzato, V. A., Rosenke, R., Hanley, P. W., Saturday, G., Scott, D., Fischer, E. R., & de Wit, E. (2020). *Respiratory disease and virus shedding in rhesus macaques inoculated with SARS-CoV-2*. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/2020.03.21.001628

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

Subiaul, F., Cantlon, J. F., Holloway, R. L., & Terrace, H. S. (2004). Cognitive imitation in rhesus macaques. *PsycEXTRA Dataset*. https://doi.org/10.1037/e604012013-026

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z