DSC 480 Midterm
Xuyang Ji, Mimidoo Gyoh, Angelo Kelvakis

I. Introduction

Network Analysis Corp (NAC) was tasked with compiling company-level network data on the employees at Real Shade Sunglasses company (RSS). Members of the NAC data science team interviewed employees from each department of RSS: Executive, Marketing, Sales, HR, Distribution, Manufacturing, and Finance. The employees that were interviewed were asked basic questions about everyday life at their company in an effort to compile three different networks. The networks that were developed were based on the three research questions:

1. Which coworkers do you notice spending time together during lunch break or outside of work?
2. Who do you observe your coworkers interacting with each other on work-related matters?
3. Who do you observe your coworkers giving advice to?

From those three questions we were able to generate three networks: Friendship, Work, and Advice. The Friendship Network houses information about employees who talked about hobbies, discussed their personal lives, and had informal discussions about work-related topics in a non-professional capacity. The Work Network housed all of the data on who was described working together on projects, giving updates on material, and engaged in meetings. The Advice Network has information on who gives professional advice to each other and can give insights into the hierarchy of the company and who feels like they need information and who gives it. The Advice Network also includes data from a mentorship program where employees were paired with each other based on seniority and a pre-filled personal survey in order to promote cross-department cohesion within the company. This data was added on top of the research question data.

II. Scope

The data was collected in an interview style with 3 members of NAC interviewing one employee at a time and asking the three research questions above in a randomized pattern with randomized selection of the employees so that no single department was sampled more than another. The response from the interviewee was recorded in third person with the names of the employees recorded and then encoded in the final data. Along with the research questions and interviews, a post-interview survey was sent out asking everyone who was interviewed about their physical appearance (hair color), gender, corrective lenses usage, and hobbies. This data was stored within the code book where survey questions were joined to company data on employee names, age, years worked, and department. You can find keys to each numeric attribute within this codebook as well.

The three networks were compiled from edgelists where question data was converted into interaction data for the edgelist. This was then formatted using R to compile adjacency matrices for analysis. The R packages “xUCINET”, “sna”, and “igraph”, were used to collect statistical data on the networks in order to compare metrics like density, transitivity, dyadic and triadic relationships, and centrality. The networks were also visualized and information on how the network related to the information from the research questions was analyzed.

DSC 480 Midterm
Xuyang Ji, Mimidoo Gyoh, Angelo Kelvakis

III. Governance

The Real Shade Sunglasses company is headed by a team of executives a1, a2, a3, and a4. These four participants guide the daily work of all the departments and make sure that the Shade is indeed real. They are supported by senior staff of each department and were all included in the interview process. At this company, there are contracts and agreements in place. Some data may be restricted and some data will need to be encrypted and anonymized. A mentorship program was set up to ensure that all members of the departments have access to information and can better form interpersonal relationships with their colleagues at the organization. The mentorship is structured to have a lead mentor from each department and several employees from various departments. NAC was given access to some select documents by the executive teams including the mentorship roster , basic demographics and employee engagement programs (hobbies). Overall, while the hierarchy can be clearly distinguished by seniority, based on the data below, you will find that this network is mostly sparse, with weak interpersonal relationships, stronger working relationships, and contains mild advice relationships. The governance in this study is that each employee was asked by a governing participant (executive department and the heads of their departments) to talk to NAC. They were given a 10 minute time allotment and 70 members complied.

IV. Network Data Cleaning & Preprocessing

The dataset comes from the week-long interviews of 70 employees cross Executive, Marketing, Sales, Human Resources, Distribution, Manufacturing, and Finance departments in the sunglass company, in which we ask all employees to list out the relevant employee interactions in certain scenarios, the location where it happened, and the names of involved employees. Employees' personal information was collected and encoded as below, including gender, need for corrective lenses, hair color, seniority level(years spent working for the company), hobbies, and mentorship status. Following negotiations with management, the names are encoded and anonymous as a combination of letter and sequence. NAC coded each partner with the name a# where # represented a numeric increment which serves as the identification for the participants in this study.

Department	Code	Gender	Code	Lenses	Code	Hair	Code	Seniority	Code	Hobby	Code	Mentorship	Code
Executive	0	Female	0	No need	0	Blonde	0	Entry: 0-6	0	Poker	0	Mentor	1
Marketing	1	Male	1	Needs	1	Brown	1	Junior: 6-10	1	Tennis	1	Mentee	0
Sales	2					Black	2	Senior: 10+	2	Cooking	2		
Human Resources	3					Red	3	*yrs have worked		Camping	3		
Distribution	4									Golf	4		
Manufacture	5									Coffee Roasting	5		
Finance	6									Video Games	6		

DSC 480 Midterm
Xuyang Ji, Mimidoo Gyoh, Angelo Kelvakis

To better understand the relationships within the company, we have observed various types of interactions, including mentorship, friendship, work-related, and etc. Our dataset is stored as comp.network where all friendship, working, and advice ties are in edgelist format in csv file. In the undirected working and friendship edgelist, a value of 1 indicates there is a mutual relationship presents between the vertices; whereas the a value range from 1 to 4 in the directed and weighted advices network indicates the quality and importance of the advices given from the source vertex to the receiver based on their seniority level. To ensure a smooth conversion from the edgelist to adjacency matrix, all possible vertices' names are appended with a weight value of 0; then sorted the rows into appropriate sequences.

Meantime, given the fact there were intragroup interactions within the departments and intergroups interactions between cross-functional departments; we first use matrix multiplication to get the adjacency matrix of the intragroup interactions with the mmult formula in Excel, and then use R to transform the intergroup interactions which stored as a edgelist into an adjacency matrix. Finally, these two matrices are merged with the condition that if two vertices have a relationship present in the intergroup adjacency matrix, then fill the value of 1; otherwise, fill the value as in the intragroup adjacency matrix.

**Details about the formula can be found in Merged_working workbook.*

After importing the relevant CSV file into R, additional rows are firstly trimmed to keep data consistent and clean. Finally igraph objects are created for each network with appropriate flags, such as directed and weighted in the 'advices' network. The graph.data.frame function would simultaneously import the vertex attributes and edge attributes if there are more than 2 columns. An important step is to filter out any edges with a weight of 0, where *simplify* is used. Due to the way the data is imported, there are self-loops with a weight of 0 along the diagonal of the matrix, which would cause problems for sociomatrix transformation. To retain multiple edges while removing loops, the *remove.loops* parameter is set to true. Finally, the edgelist matrices are transformed into adjacency/sociomatrix with *get.adjacency.* For the sociomatrix of 'advices', the attribute weights is added simultaneously by setting *attr= 'weights*'.

```r
library(igraph)
friendship <- read.csv('EdgeList_ds/friendship_edgelist.csv')
friends <- graph.data.frame(as.matrix(friendship), directed = FALSE)
friends2 <- simplify(friends, remove.multiple = FALSE, remove.loops = TRUE)
friend_adjacency <- get.adjacency(friends2, sparse = FALSE)
class(friend_adjacency)
friend_adjacency
write.csv(friend_adjacency,"friend_adjacency.csv")

working <- read.csv('EdgeList_ds/working_edgeList.csv', header=FALSE)
work <- graph.data.frame(as.matrix(working), directed = FALSE)
work2 <- simplify(work, remove.multiple = TRUE, remove.loops = TRUE)
work_adjacency <- get.adjacency(work2, sparse = FALSE)
    You, 21 hours ago • uploaded working_adjacency matrix …
work_adjacency
write.csv(work_adjacency,"work_adjacency.csv")

advices <- read.csv('EdgeList_ds/advice-Edgelist.csv', header = TRUE)
advices <- advices[c(1:135),]
advices <- graph_from_data_frame(advices, directed = TRUE)
advices2<- simplify(advices, remove.multiple = TRUE, remove.loops = TRUE)
advice_adjacency <- get.adjacency(advices2, sparse = FALSE, attr = 'weight' )
write.csv(advice_adjacency,"advice_adjacency.csv")
```

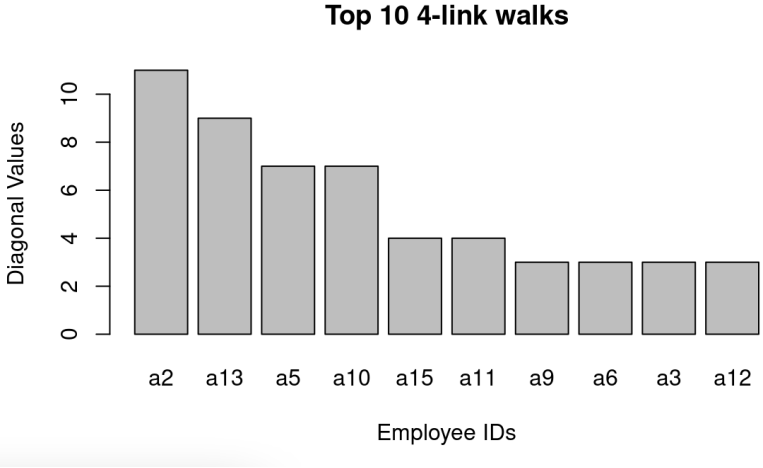
DSC 480 Midterm
Xuyang Ji, Mimidoo Gyoh, Angelo Kelvakis

V. Data Analysis

Friendship Network

I. I. Network Shape And Basic Descriptions

One method of calculating who is the most well connected within a friendship network, the adjacency matrix was multiplied 4 times to see which employees had paths back to themselves. As seen by the graph of the first 10 sorted values, a2 has 11 different 4-vertex paths to get back to themselves, a13 has 9, a5 and a10 have 7, a15 and a11 have 4, and several others have 3, 2, and 1. The conclusion can be drawn that a2 and a11 are the most popular employees at the sunglasses company. They have the highest amount of connections in a friendship network and have been observed to have the most performing social interactions. a5 and a10 are close behind them with slightly fewer. Only the top ten are shown as trailing values are all equal to 1 or 0.



In order to gather metrics on this network, density is compared to the transitivity and the number of isolates. In this network there are 70 nodes. This network is very sparse as the density is only 0.0083. This means that the network is very spread out with many more nodes than links and people who are friends only happen between a handful of individuals. There are no wide-spread networks of friendships. The Transitivity of this network is equal to 0.3 meaning that there are more triads than dyads. This can be interpreted as the people who have friendships are connected to another employee and are not simply one on one. With almost half of the network containing isolates, it may be concluded that either the interviewers did not collect data on the actual friendships present in the company, or this is a business culture where friendships are not openly displayed / not present.

II. Measure Of Centrality

The three methods of centrality used to compare the network are degree, betweenness, and eigenvector. As seen by the ordered outputs of the centralized data, the outcome from the matrix multiplication is similar with the degree mapping to the order of the 4-connection output from above. The between data starts to diminish quickly after the first couple entries and the EV values follow somewhat of a similar pattern as the degree values. Our data does not show a strong correlation between any of the metrics where a change in degree explains only 65% of the variation in between and only 52% of variation in ev. The correlation between ev and between is even lower with only 28% explained.

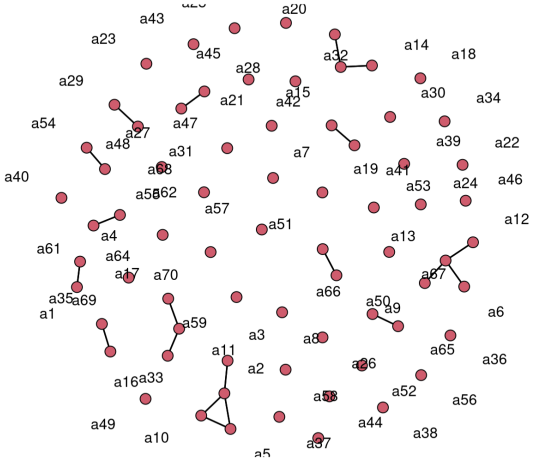
	degree	between	ev
a2	3	2	1.000000e+00
a13	3	3	0.000000e+00
a5	2	0	8.546377e-01
a10	2	0	8.546377e-01
a11	2	1	2.722788e-16
a15	2	1	9.075960e-17
a1	1	0	6.806970e-17
a3	1	0	4.608111e-01
a4	1	0	1.134495e-16
a6	1	0	3.630384e-16
a9	1	0	3.630384e-16
a12	1	0	3.403485e-16
a14	1	0	2.268990e-16
a16	1	0	2.949687e-16
a17	1	0	2.722788e-16
a20	1	0	1.815192e-16
a27	1	0	4.537980e-17

…

	degree	between	ev
degree	1.0000000	0.6508570	0.5228337
between	0.6508570	1.0000000	0.2798262
ev	0.5228337	0.2798262	1.0000000

III. Graphing Of The Network

The numeric outputs from calculating transitivity and density can be more clearly seen with the graph output where the sparsity is easy to see and the lack of complex networks shows why the density is so low. This is again shown through the histogram of geodesics where the lengths of all the connections are zero, meaning that the distance between employees cannot be calculated due to the lack of linkages between nodes.



DSC 480 Midterm
Xuyang Ji, Mimidoo Gyoh, Angelo Kelvakis

Working Network

I. Clustering coefficient and Basic Description

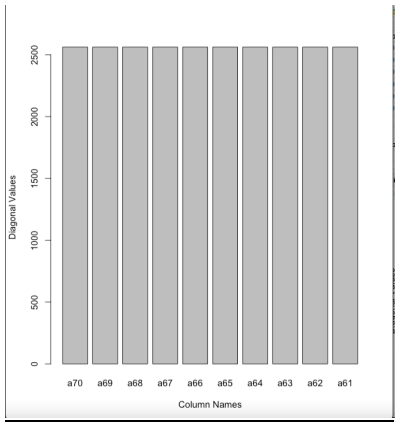
In this network there are 70 nodes. The network is relatively dense with a density of 0.1602 suggesting that the network is more dense than sparse. There are closely connected networks of work. It has a clustering coefficient (transitivity) of 96% meaning this network has a high level of transitivity. This means individuals are more likely to have connections to other individuals who are also connected to each other therefore forming clusters within the network. This can have important implications for diffusion of information.

There are no isolates. It may be concluded that every participant in the working network has at least one connection.

II. Network Paths (4-Link path)

One method of calculating who is the most well connected within a working network, the adjacency matrix was multiplied 4 times to see which employees had paths back to themselves.

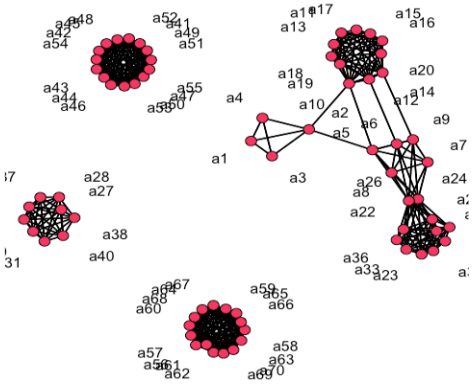
As seen by the graph of the last ten sorted values, they all have about 2600 4-vertex paths to get to themselves. This means that they have a similar and high level of connectivity within a work setting in the sunglass company. They all have a high amount of connections within a working network.

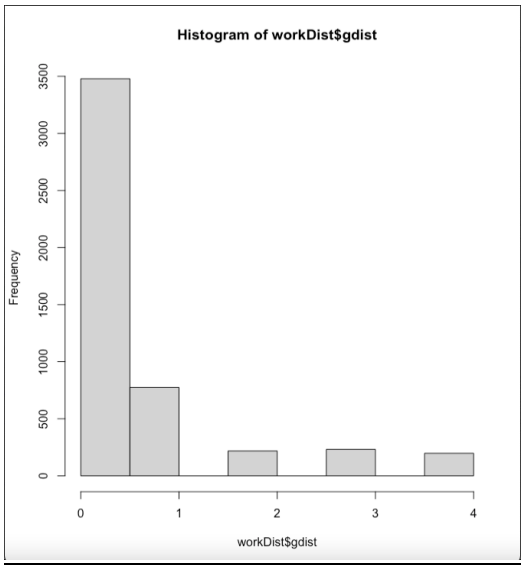


III. Graph Of The Network

The numeric output from calculating the clustering coefficient (transitivity) and density can be seen clearly with the graph output. This is a left skewed distribution. This suggests that there are many pairs of nodes with relatively small geodesic distances but only a few pair of nodes with large geodesic distances. It also suggests that these nodes are closely connected and that information or influence can flow easily between them.

DSC 480 Midterm
Xuyang Ji, Mimidoo Gyoh, Angelo Kelvakis





Advising Network

I. Data Transformation - Dichotomization

While the weights in the directed advising network indicates the importance and quality of the advices and mentorship had given by the source vertices, with 1 being very little and 5 being a great deal. Dichotomization technique is used to transform the asymmetric raw data, which refers to converting valued data into binary data. The reason behind this is that graph-theoretic methods are only applicable to binary data. We take the valued adjacency matrix and set all cells with a tie strength greater than the threshold value of 0 as 1, ,and set al the remaining cells to 0. After the transformation, there are approximately 2.4% of the vertice pairs have a tie (i.e. density=0.024), with a transitivity score of 0.26 and 4830 possible dyads. The 26% transitivity score shows the ratio of connected triples existed in the network, further indicating the company's top-down management/advising system.

II. Component Size & Distribution

Within the advising relationship, one initial assumption is that coworkers with higher socially significant attributes are more likely to mentor/advise those with lower attribute values, such as seniority, age, and status. For seniority, we construct employees with the same seniority level as a sociomatrix and convert into a relative seniority difference matrix. For age, we might use absolute difference in age; however, considering the diversity and company policy, we decided to focus on seniority level only. Looking at the distribution of the directed advising network, the false output for is_connected command indicates it does not have a directed path from each vertex to all other vertices, hence our network is rather sparse.

DSC 480 Midterm
Xuyang Ji, Mimidoo Gyoh, Angelo Kelvakis

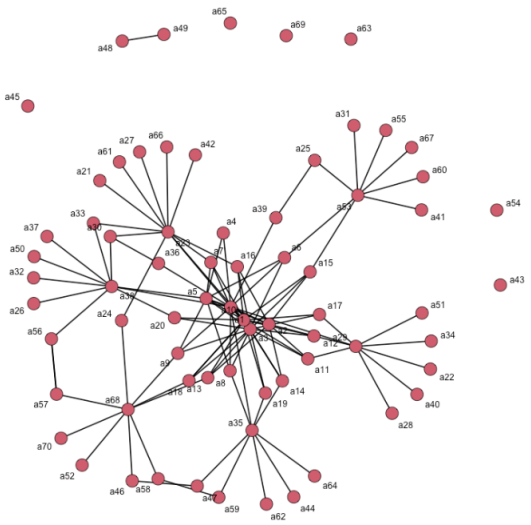
As listed below, the membership indicates the cluster id to which each vertex belongs, whereas most employees are advice-seekers as in cluster 1 with 62 employees while a few other coworkers with higher seniority level are mentor represented in higher vector. Looking at the plotted largest component on the right, it can be inferred that most employees can receive advice/mentorship from coworkers with higher seniority level, either directly or indirectly, since the “weak” attribute also considers semi-paths. Upon investigation, there are 6 isolated employees who are enabled to get mentorship from others. On the other hand, based on the closeness centrality measures, employees including a48, a49, a1, and a10 require most steps to get connected to. In conclusion, while most employees are well-connected in the mentorship/advising system, some entry level employees might be afraid to reach out for help. Although the CEO of the company seems hard to get connected with at the first glance, she is closely connected with other executives and department leads, as shown in the *Figure.Advice Network Graph* below.

```
> library(igraph)
> ##convert into graph object
> adviceGraph <- graph_from_adjacency_matrix(sunglass_network$Dichotimized_advice,
+                                            mode = 'directed',
+                                            weighted = TRUE,
+                                            diag = FALSE)
> components(adviceGraph, mode = c('weak'))
$membership
 a1  a2  a3  a4  a5  a6  a7  a8  a9 a10 a11 a12 a13 a14 a15 a16 a17 a18 a19 a20
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
a21 a22 a23 a24 a25 a26 a27 a28 a29 a30 a31 a32 a33 a34 a35 a36 a37 a38 a39 a40
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
a41 a42 a43 a44 a45 a46 a47 a48 a49 a50 a51 a52 a53 a54 a55 a56 a57 a58 a59 a60
  1   1   2   1   3   1   1   4   4   1   1   1   1   5   1   1   1   1   1   1
a61 a62 a63 a64 a65 a66 a67 a68 a69 a70
  1   1   6   1   7   1   1   1   8   1

$csize
[1] 62  1  1  2  1  1  1  1

$no
[1] 8

> lgc<- component.largest(adviceDN, connected="weak",
+                         result = 'graph',
+                         return.as.edgelist = FALSE)
> gplot(adviceDN,vertex.col=2+lgc)  #Plot with component membership
> #Plot largest component itself
> isolates(adviceDN)
[1] 43 45 54 63 65 69
> closeness(adviceGraph, mode='all')%>% sort(decreasing = TRUE)%>%.[1:4]
       a48         a49          a1         a10
1.000000000 1.000000000 0.008547009 0.007518797
```



The *Figure.Advice Network Graph* uses the Frucheterman-Reingold method models, representing the vertices as a collection of magnets and springs. As shown, employees with ID number from a1 to a5 are the executive/management team within the sunglass company, hence often play a role as advice-giver to department leads, while employees with higher ID number are most likely to be the advice-seeker, receiving mentorship/advices from their respective department leads.

To classify dyads in the directed graph, the relationships between each pair of vertices is measured, across three states: mutual, asymmetric or non-existent. Based on the output, there are 2 pairs with mutual connections, 112 pairs with asymmetric connections, and 2301 pairs with no connection between them. Hence, it could be concluded that mentorship and the advising systems are extremely exclusive with employees across various functional departments, where often the department leads give high-quality feedback and mentorships to employees within the same teams.

DSC 480 Midterm
Xuyang Ji, Mimidoo Gyoh, Angelo Kelvakis

```
> ## Vignette #4 Dyads and Triags
> dyad.census(adviceGraph)
$mut
[1] 2

$asym
[1] 112

$null
[1] 2301

> triad.census(adviceGraph)
 [1] 47708  6234   119   466    55   101     7     9    40     0     0     1
[13]     0     0     0     0
```