Network Data Analysis - Assignment 1

Felipe M. Megale, 100806980

February 2022

Contents

1 Pre-processing data

To complete the pre-processing of the original dataset, I used Pandas and Numpy. The first step for pre-processing the original file, after downloading it, was to remove all rows that had issues. Specifically speaking, this issue is not having either person name or ID. If either of these two information are missing, the row is removed from the original set. After removing those rows, I also removed those columns that did not have a label i.e., a question, and the two last questions that were further apart from the first 7. Furthermore, in order to have a more intelligible data set, I renamed the first column from *"Unnamed: 0"* to *"Name"*. Then, I isolated all questions and for each one of them, I created a new data set comprised of three attributes: *"Name"*, *"ID"*, *"Question"*. This resulted in 7 new data sets, one for each question, which were saved as CSV files. The questions are:

1. "Which person you have hear of their voice or seen their faces?"

2. "Which person you have met (in person+online) and exchange conversation?"

3. "Which person you have collaborated with?"

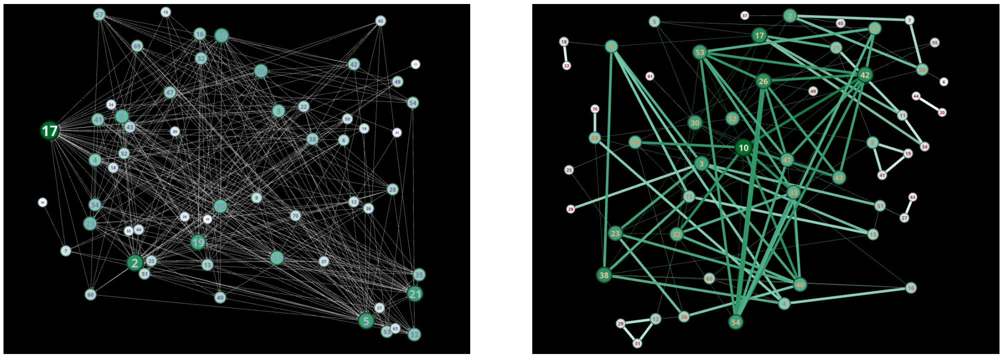
4. "Which person you have eye contact?"

5. "Which peson you have eaten lunch with?"

6. "Which person you have shared a ride?"

7. "Which person you have taken at least two courses with?"

After having separated the questions, I processed the inputs of each question in each individual file. The reason for this was to simply adequate each person's input to comma-separated IDs by removing spaces and trailing characters such as commas and hyphens. Finally, for each of the 7 questions, I created 7 other files in the CSV format Gephi expects. For example, a line with the following values *"1,2,3,4,5"* means that node 1 is connected to nodes 2, 3, 4 and 5.



(a) Network 1

(b) Network 2

Figure 1: Networks 1 & 2

2 Analyses of 6 networks

2.1 Available networks

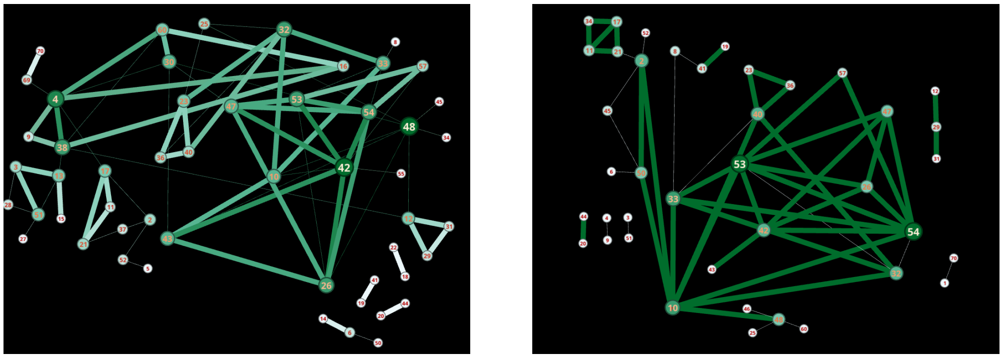
For this assignment, out of the 7 available networks, the ones I chose to work with are networks 1 through 5, and 7.

Before describing each network individually, one characteristic all of them share is that the edges in all 7 graphs are unweighted for the same reason. The questions asked *"which people?"*, meaning that for each person in the list, you only place an edge if the answer is yes. Because it is binary (edge or no edge), all graphs have unweighted edges. However, if we had a question asking *"how many times?"*, then there is a possibility of having weighted edges. Also, Gephi does not render a node if its degree equals zero.

2.1.1 Network 1

The first network asks the following question: *"Which person you have hear of their voice or seen their faces?"*.

Directed or undirected? This network is an example of a directed graph. The reason for this is that in online conferences, it is not all participants who speak and/or have their computer cameras on. Therefore, I may see other people's faces and/or hear their voices, but they may not see or hear me.



(a) Network 3

(b) Network 4

Figure 2: Networks 3 & 4

2.1.2 Network 2

The second network asks the following question: *"Which person you have met (in person+online) and exchange conversation?"*.

Directed or undirected? This network is an example of an undirected graph. In order to have a conversation, both parties must engage. If only one of them speak, there is no dialog. Therefore, is is not possible to exchange conversation.

2.1.3 Network 3

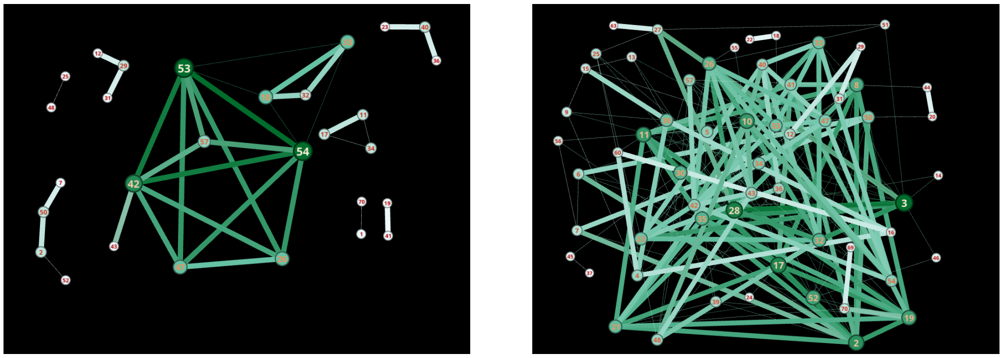
The third network asks the following question: *"Which person you have collaborated with?"*.

Directed or undirected? This is an undirected graph because collaboration must be enforced by both parties. It has to be a mutual agreement.

2.1.4 Network 4

The fourth network asks the following question: *"Which person you have eye contact?"*.

Directed or undirected? This network is undirected because in order to establish eye contact, two people must engage. There is no way for one person to look into someone else's eyes and not be looked back.



(a) Network 5

(b) Network 7

Figure 3: Networks 5 & 7

2.1.5 Network 5

The fifth network asks the following question: *"Which peson you have eaten lunch with?"*.

Directed or undirected? This is an undirected graph because eating lunch with another person implies that both people had to simultaneously engage in the activity.

2.1.6 Network 7

The seventh network asks the following question: *"Which person you have taken at least two courses with?"*.

Directed or undirected? This network is an undirected graph because you cannot be simultaneously enrolled in a course another person has not and consider that as taking a course together. Also, the course must be taken in the same year.

Metric	**Net. 1**	**Net. 2**	**Net. 3**	**Net. 4**	**Net. 5**	**Net. 7**
Num. Nodes	60	60	59	60	59	60
Num. Edges	426	120	73	48	30	214
Density	0.120	0.067	0.042	0.027	0.017	0.120
Avg. Clust. Coef.	0.394	0.304	0.296	0.241	0.168	0.488
Num. Nodes SCC	49	-	-	-	-	-
Num. Nodes WCC	58	-	-	-	-	-
Num. Nodes CC	-	52	42	20	10	59
Avg. Path Len. SCC	2.599	-	-	-	-	-
Avg. Path Len. CC	-	3.143	4.234	2.857	1.711	2.654
Diameter of SCC	7	-	-	-	-	-
Diameter of CC	-	7	11	6	4	6

Table 1: Network statistics

3 Metrics of 6 networks

In this section I will present some statistics about the 6 chosen networks. The results will be presented in tables. All metrics and plots were calculated and generated using the Python package NetworkX.

3.1 Number of nodes

The number of nodes of each graph is the amount of people who have participated in the survey. Whether they are connected to someone else, or not, they will be represented as nodes.

3.2 Number of edges

The number of edges of each graph will be the amount of connections that exist between each person for each question.

3.3 Edge density

Graph density tells us how connected nodes are between each other. If the density value is high, we say the graph is connected, if the density value is low, we say it is sparse. For undirected graphs, this metric can be calculated as

$$D_{undirected} = \frac{2|E|}{|N|(|N|-1)} \tag{1}$$

and the density for directed graphs is defined as

$$D_{directed} = \frac{|E|}{|N|(|N|-1)} \tag{2}$$

where E is the number of edges and V is the number of nodes in the graph.

3.4 Degree distribution

The degree distribution can superficially tell us what are the preferences for people when connecting to each other. Figure 4 plots the degree distributions for the 6 chosen networks.

3.5 Average clustering coefficient

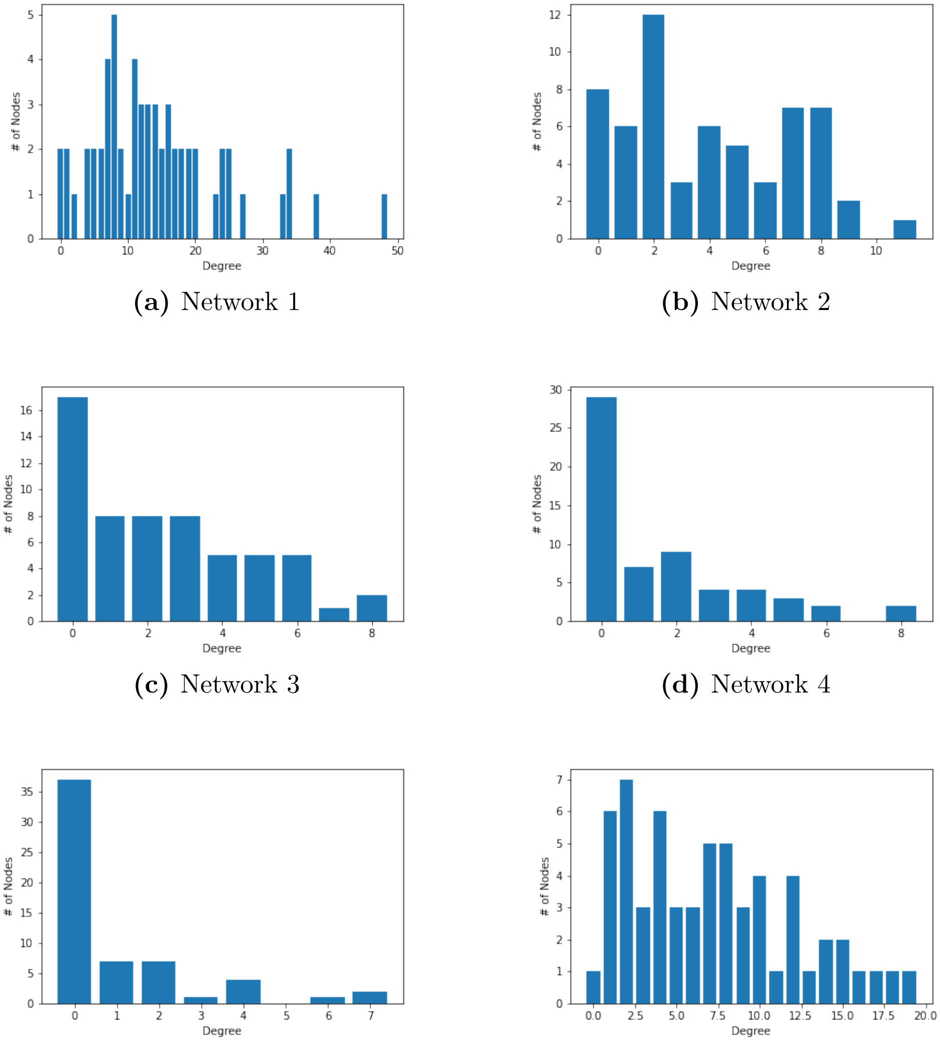
The average clustering coefficient for a graph helps determine how transitive a relationship is. For example, if persons A and B are friends, and persons A and C are friends, there is a high chance that persons B and C are also friends. The clustering coefficient is defined as

$$C_i = \frac{2e_i}{k_i(k_i-1)} \tag{3}$$

where e_i is the number of edges between the neighbors of node i.

The average clustering coefficient of the graph is calculated as

$$\langle C \rangle = \frac{1}{N}\sum_{i}^{N} C_i \tag{4}$$



(a) Network 1

(b) Network 2

(c) Network 3

(d) Network 4

(e) Network 5

(f) Network 7

Figure 4: Degree distributions

where N is the number of nodes in the graph, and C_i is the clustering coefficient of node i.

3.6 Number of nodes in strongly connected component (SCC)

The strongly connected component (SCC) metric can only be obtained from directed graphs. Since only the first network is directed, it is the only one that can provide this value. For networks 2 through 5, and 7, the values are from the connected components. Refer to table 1 for the values.

3.7 Number of nodes in weakly connected component (WCC)

The weakly connected component (WCC) metric can only be obtained from directed graphs. Since only the first network is directed, it is the only one that can provide this value. For networks 2 through 5, and 7, the values are from the connected components. Refer to table 1 for the values.

3.8 Average path length in SCC

The average path length metric indicates how far apart two nodes are from each other in the connected graph. In other words, how many jumps, in average, it takes to reach other nodes. For a directed graph, it is calculated as

$$\langle d \rangle \equiv \frac{1}{2L_{max}} \sum_{i,j \neq i} d_{ij} \tag{5}$$

whereas for undirected graphs, it is calculated as

$$\langle d \rangle \equiv \frac{1}{L_{max}} \sum_{i,j > i} d_{ij} \tag{6}$$

For networks 2 through 5, and 7, I calculated the average path length in the connected component because these networks are undirected graphs, thus not being possible to determine strongly connected components.

3.9 Diameter of SCC

This metric represents the maximum shortest distance between two nodes in a connected graph. It is be represented as

$$diameter \equiv \max_{ij} d_{ij} \tag{7}$$

For the first network, I collected the diameter of the strongly connected component. However, since all other networks are undirected graphs, I collected the diameter of the largest connected component.

3.10 Community detection

To detect communities in each of the chosen networks, I ran the Girvan–Newman algorithm, implemented in NetworkX. Figure 5 illustrates the communities in each network. The communities the algorithm found make sense, given that it ran by removing the edges with highest betweenness, separating the communities the edge held together. Also, by comparing with figures 1, 2, and 3, we can see that the nodes with most connections between each other form a community the algorithm was able to find. Another interesting feature that the community detection algorithm allows us to perceive is that the more sparse the graph, the more communities we have. Figures 5c, 5d, and 5e depict this behavior.

3.11 Centrality Measures

Centrality tries to determine which node is the most central in a graph. The four centrality measures will be used to determine which node is the most important in each network. The measures are in-degree, out-degree, betweenness, and closeness. The choice for each centrality measure for each graph was arbitrary.

3.11.1 Network 1

For this network, in-degree and out-degree will be used to determine the two most central nodes. Because it is a directed graph, we can use these measures. The in-degree centrality of a node is calculated based on how many edges arrive in it. Similarly, the out-degree centrality of a node is calculated based on how many edges leave it. Table 2 summarizes the two most people.

Rank	Person	Score
1	17	0.576
2	5	0.508

(a) In-degree

Rank	Person	Score
1	10	0.271
2	19	0.254

(b) Out-degree

Table 2: Network 1 centrality

Rank	Person	Score
1	10	0.186
2	42	0.152

(a) Degree

Rank	Person	Score
1	17	0.169
2	30	0.131

(b) Betweenness

Table 3: Network 2 centrality

3.11.2 Network 2

The centrality measures analyzed for this network were degree centrality and betweenness centrality. Degree centrality is similar to in-degree and out-degree centrality measures, except for the direction of the edges, which do not exist. The degree centrality measure for a node is defined by the amount of edges connected to it. Betweenness, on the other hand, is defined by the amount of shortest paths that go through a given node, and is calculated as follows

$$C_B(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}} \tag{8}$$

where g_{jk} is the amount of shortest paths connecting nodes j and k, and $g_{jk}(i)$ is the node currently being analyzed. Table 3 summarizes the values found.

3.11.3 Network 3

The chosen centrality measures for this network were degree and closeness. Degree centrality has been previously defined. However, the closeness centrality measure is defined by the average length of shortest paths between a

Rank	Person	Score
1	48	0.137
2	42	0.137

(a) Degree

Rank	Person	Score
1	48	0.245
2	30	0.243

(b) Closeness

Table 4: Network 3 centrality

Rank	Person	Score
1	10	0.039
2	33	0.032

(a) Betweenness

Rank	Person	Score
1	10	0.169
2	33	0.165

(b) Closeness

Table 5: Network 4 centrality

node and all other nodes in a graph. It can be calculated as

$$C_C(i) = \left[\frac{1}{N-1} \sum_{j=1}^{N} d(i,j) \right]^{-1} \tag{9}$$

where N is the number of nodes in a graph, and $d(i,j)$ is the distance between nodes i and j. Table 4 summarizes the centrality results for this network.

3.11.4 Network 4

The centrality measures chosen to analyze from this network were betweenness and closeness. These two measures have been previously defined. Table 5 summarizes the values found.

3.11.5 Network 5

The centrality measures analyzed for this network were degree and betweenness centralities. These measures already have been introduced. Table 6 summarizes the two most influential nodes of this network and their centrality score.

Rank	Person	Score
1	53	0.120
2	54	0.120

(a) Degree

Rank	Person	Score
1	42	0.005
2	53	0.004

(b) Betweenness

Table 6: Network 5 centrality

Rank	Person	Score
1	3	0.322
2	28	0.305

(a) Degree

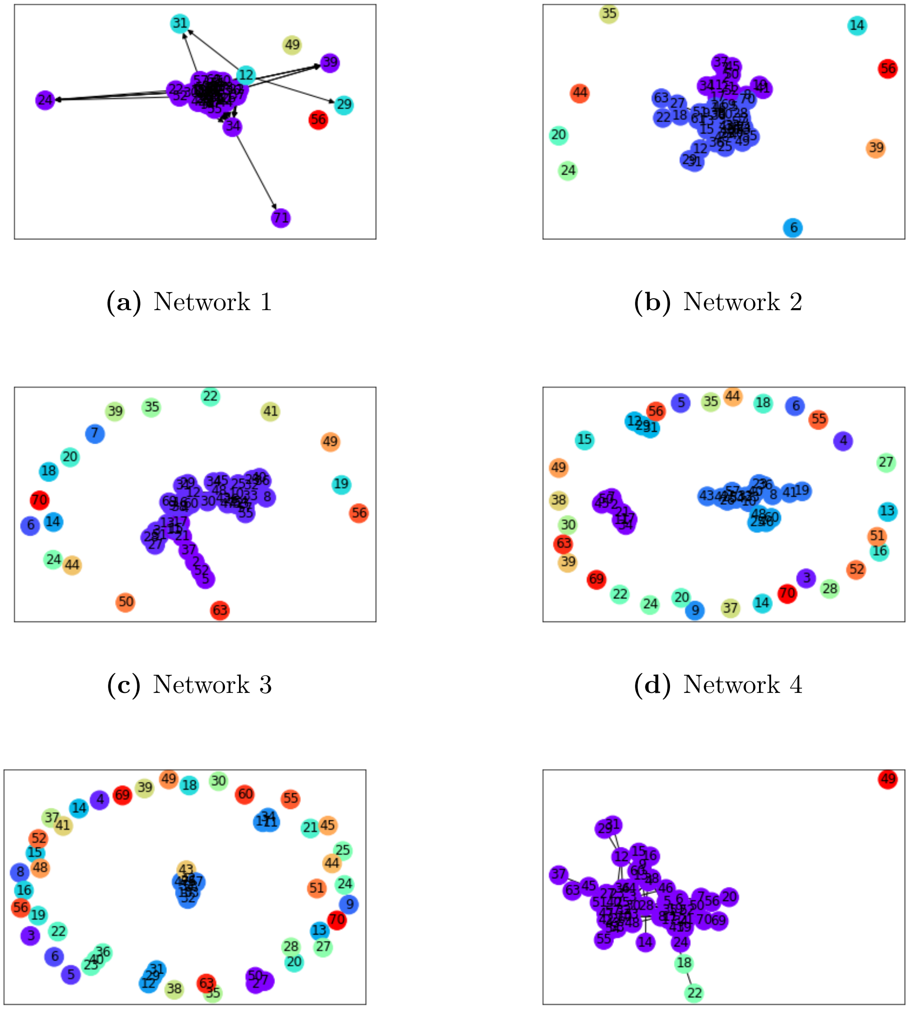
Rank	Person	Score
1	28	0.543
2	17	0.509

(b) Closeness

Table 7: Network 7 centrality

3.11.6 Network 7

The chosen centrality measures for network 7 were degree and closeness. I will refrain from going deeper into these concepts since they have already been introduced. Table 7 summarizes the centrality findings for this network.



(a) Network 1

(b) Network 2

(c) Network 3

(d) Network 4

(e) Network 5

(f) Network 7

Figure 5: Detected Communities

4 Insights

All the metrics previously calculated and visualized allow us to extract insightful knowledge about the networks that exist in our graduate program. Tackling the aspects of low degree distribution, low density and high number of communities, we can see that the questions *"Which person you have collaborated with?"*, *"Which person you have eye contact?"*, and *"Which peson you have eaten lunch with?"* create the most sparse graphs. That may be due to the fact that most classes are still online and people haven't gotten the chance to interact with each other in a more meaningful way.

It is also interesting to notice unexpected behaviors on sparse networks. Let us take network 3 as example. It was one the three networks that were low in density and very sparse. However, the existing connected component was well connected. This means that despite existing many people who did not collaborate with their peers, the ones who did did extensively.

Finally, discovering which nodes in each graph are more central is important. The reason for this is that knowing which people connect two or more components may contribute to how the people in these networks interact and collaborate with each other. Removing some of the key