In [1]:
```
# Importing the required libraries
import pandas as pd
import networkx as nx
import matplotlib.pyplot as plt
G=nx.Graph()

from datetime import date
today = date.today()

print(today)
```

```
2021-06-09
```

CUNY SPS DATA 620

WEEK 2 ASSIGNMENT 1

This assignment is about the Enron Email dataset, the dataset contains > 370K edges, we will work with various subsets of those edges.

The dataset was sourced from https://snap.stanford.edu/data/email-Enron.html

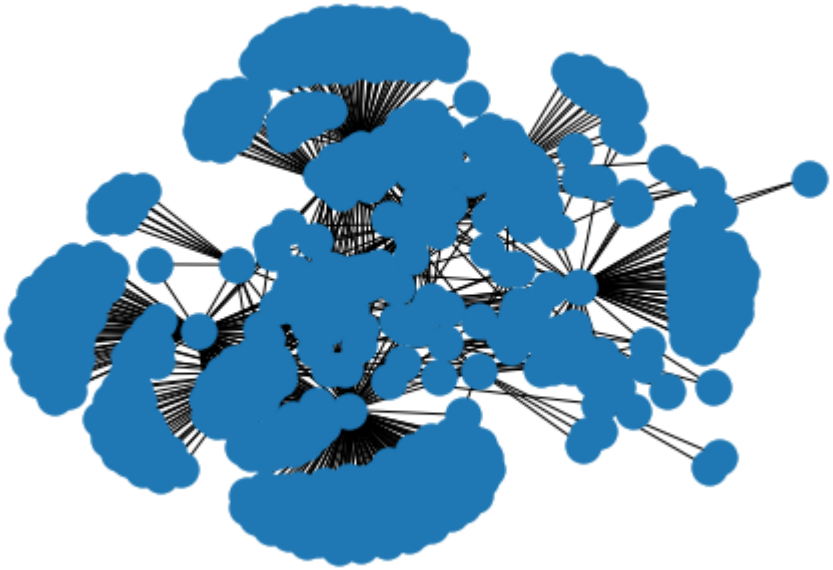
This work is a group effort, the group members are Ramnivas Singh, Deepak Sharma, Tage Singh,

The cell below provide an overview of **1,000** records of the ENRON Email dataset

In [2]:
```
data_frame_1k = pd.read_csv('https://raw.githubusercontent.com/tagensingh/DATA620-W2-A1/main/gephi_import_1000_records.cs

data_frame_1k.head()

G=nx.from_pandas_edgelist(data_frame_1k, 'Source', 'Target',edge_attr=None)

#G = nx.read_edgelist("enron_1000_records.txt") # selecting the subset of 1000 records

print(nx.info(G))

nx.draw(G)

plt.show()

print('The network density of 1000 nodes is :',nx.density(G) )# Displaying the density metric of the datset of 1000 email
```

```
print(' The number of NODES in this dataset is :',nx.number_of_nodes(G)) # Displaying the total number of nodes

print(' The number of EDGES in this dataset is :',nx.number_of_edges(G)) # Displaying the total number of edges

# nx.degree_centrality(G) - This produce a list of the degrees of centrality for each node.-The code runs quickly but the

print(' The RADIUS of this dataset is :',nx.radius(G)) # Displaying the radius metric

print(' The DIAMETER of the dataset is : ',nx.diameter(G)) # Displaying the diameter of the dataset
```

```
Name:
Type: Graph
Number of nodes: 528
Number of edges: 910
Average degree:   3.4470
```



```
The network density of 1000 nodes is : 0.006540739462940601
 The number of NODES in this dataset is : 528
 The number of EDGES in this dataset is : 910
 The RADIUS of this dataset is : 2
 The DIAMETER of the dataset is :  4
```

The cell below provide an overview of **100,000** records of the ENRON Email dataset

This subset of the data is too large to generate a graph using current hardware config

In [3]:

```
#G = nx.read_edgelist("email-enron-working-all-records.txt")  -
```

```
#This code wil read all 367K records, this will take significant time to compute!


data_frame_100k = pd.read_csv('https://raw.githubusercontent.com/tagensingh/DATA620-W2-A1/main/gephi_import_100000_record

data_frame_100k.head()

G=nx.from_pandas_edgelist(data_frame_100k, 'Source', 'Target',edge_attr=None)

#G = nx.read_edgelist("enron_100000_records.txt") # selecting the subset of 1000 records

print(nx.info(G))

# nx.draw(G)  - We cannot run this for the 100,000 records since the compute time is significantly extended

# plt.show()

print('The NETWORK DENSITY of 100000 nodes is :',nx.density(G) )# Displaying the density metric of the datset of 1000 ema

print(' The number of NODES in this dataset is :',nx.number_of_nodes(G)) # Displaying the total number of nodes

print(' The number od EDGES in this dataset is :',nx.number_of_edges(G)) # Displaying the total number of edges

# nx.degree_centrality(G) - This produce a list of the degrees of centrality for each node.-The code runs quickly but the

print(' The RADIUS of this dataset is :',nx.radius(G)) # Displaying the radius metric

print(' The DIAMETER of the dataset is : ',nx.diameter(G)) # Displaying the diameter of the dataset
```

```
Name:
Type: Graph
Number of nodes: 19483
Number of edges: 81378
Average degree:   8.3537
The NETWORK DENSITY of 100000 nodes is : 0.0004287929519501978
 The number of NODES in this dataset is : 19483
 The number od EDGES in this dataset is : 81378
 The RADIUS of this dataset is : 4
 The DIAMETER of the dataset is :  7
```