Data 620 - Week 3: Assignment 2

Bikram Barua and Leticia Salazar

February 12, 2023

Overview:

This week's assignment is to:

1. Load a graph database of your choosing from a text file or other source. If you take a large network dataset from the web (such as from Stanford Large Network Dataset Collection), please feel free at this point to load just a small subset of the nodes and edges.
2. Create basic analysis on the graph, including the graph's diameter, and at least one other metric of your choosing. You may either code the functions by hand (to build your intuition and insight), or use functions in an existing package.
3. Use NetworkX to visualize the data
4. Please record a short video (~ 5 minutes), and submit a link to the video in advance of our meet-up.

Data source:

The dataset is a subset of authentication/authorization system for a web based business application. The dataset contains list of usernames with their corresponding employee Ids. The employees('Users') use the usernames as their login id as a part of the authentication process.

The web application has multiple modules, whose access is controlled using the 'Groups'. The dataset contains the list of groups and a seperate mapping of the usernames with the group names which they are granted access as a function of the authorization process for the web application access control.

Load libraries:

In [1]:
```
# data manipulation
import pandas as pd
import numpy as np

# data viz
import networkx as nx
import matplotlib.pyplot as plt
from matplotlib import rcParams
import seaborn as sns

# apply some cool styling
plt.style.use("ggplot")
rcParams['figure.figsize'] = (12, 6)
```

Load the data:

Data exploration of the data is performed below to view the size of the datasets we will be working with.

In [2]:
```
users = "https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/users.csv"
df_users = pd.read_csv(users)
```

```
print(df_users.head(5))
```

```
   username  userid
0     aauto     295
1       abc     277
2  acommins     583
3  advauto1     296
4    alexey     580
```

In [3]:
```
df_users.shape
```

Out[3]:
```
(386, 2)
```

In [4]:
```
groups = "https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/groups.csv"
df_groups = pd.read_csv(groups)
print(df_groups.head(5))
```

```
   id           group_name
0   1                ADMIN
1   2                 USER
2  15   Vehicle Visibility
3  16       DFY Operations
4  17  Document management
```

In [5]:
```
df_groups.shape
```

Out[5]:
```
(26, 2)
```

In [6]:
```
group_members = "https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/grou
df_grp_mem = pd.read_csv(group_members)
print(df_grp_mem.head(5))
```

```
     id  username  group_id  tenant_id
0  1810    smalla        38          1
1  1807    bikram        38          1
2  1809  gmspcmov        38          1
3  1811  monlucha        38          1
4  1816  krskucyk        38          1
```

In [7]:
```
df_grp_mem.shape
```

Out[7]:
```
(1000, 4)
```

Graph analysis:

Before we start to analyze the graphs, we installed packages that will allow us to work with Neo4j and Python to explore and visualize our graphs.

In [8]:
```
# toolkit for working with Neo4j from within Python applications and from the command li

#pip install py2neo
```

In [9]:
```
# operating and working with the Neo4j Graph Data Science (GDS) library.
# It enables users to write pure Python code to project graphs, run algorithms, as well

#pip install graphdatascience
```

We then make the connection with Neo4j and Python:

In [10]:
```
from py2neo import Graph, Node, Relationship

# Bikram's login info
```

```
# host = "bolt://localhost:7687/neo4j"
# user = "bikram"
# password= "bikram123"

# gds = Graph(host, auth=(user, password))


# Leticia's login info
host = "bolt://localhost:7687/neo4j"
user = "neo4j"
password= "Data620/3"

gds = Graph(host, auth=(user, password))
```

Visualization of the data:

Once the connection is established, we can then query the data sets and obtain the stats as well.

In [11]:
```
#Cypher Query

# Users csv
gds.run("LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/letisalba/Data-62
          CREATE (user: User {id: line.userid, username: line.username})").stats()
```

Out[11]:
```
{'labels_added': 386, 'nodes_created': 386, 'properties_set': 772}
```

In [12]:
```
# Groups csv
gds.run("LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/letisalba/Data-62
          CREATE (group: Group {id: line.id}) SET group.groupname=line.group_name ").sta
```

Out[12]:
```
{'labels_added': 26, 'nodes_created': 26, 'properties_set': 52}
```

In [13]:
```
# Group members csv
gds.run("LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/letisalba/Data-62
          MATCH (user: User {username: line.username}) \
          MATCH (group: Group{ id: line.group_id}) \
          CREATE (user)-[:HAS_ACCESS]->(group) RETURN user, group").to_data_frame()
```

Out[13]:

	user	group
0	{'id': '561', 'username': 'smalla'}	{'id': '38', 'groupname': 'SPECIAL_MOVE'}
1	{'id': '561', 'username': 'smalla'}	{'id': '38', 'groupname': 'SPECIAL_MOVE'}
2	{'id': '561', 'username': 'smalla'}	{'id': '38', 'groupname': 'SPECIAL_MOVE'}
3	{'id': '561', 'username': 'smalla'}	{'id': '38', 'groupname': 'SPECIAL_MOVE'}
4	{'id': '561', 'username': 'smalla'}	{'id': '38', 'groupname': 'SPECIAL_MOVE'}
...	...	...
15995	{'id': '374', 'username': 'lzfk46'}	{'id': '15', 'groupname': 'Vehicle Visibility'}
15996	{'id': '374', 'username': 'lzfk46'}	{'id': '15', 'groupname': 'Vehicle Visibility'}
15997	{'id': '374', 'username': 'lzfk46'}	{'id': '15', 'groupname': 'Vehicle Visibility'}
15998	{'id': '374', 'username': 'lzfk46'}	{'id': '15', 'groupname': 'Vehicle Visibility'}
15999	{'id': '374', 'username': 'lzfk46'}	{'id': '15', 'groupname': 'Vehicle Visibility'}

16000 rows × 2 columns

In [14]:
```
# Sub groups csv
```

```
gds.run("LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/letisalba/Data-62
         CREATE (subgroup: SubGroup {id: line.sub_group_id, subgroupname: line.sub_grou
```

Out[14]:
```
{'labels_added': 12, 'nodes_created': 12, 'properties_set': 36}
```

In [15]:
```
# CREATE RELATIONSHIP between SUB-GROUPS and Groups
gds.run(" LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/letisalba/Data-6
          MATCH (subgroup: SubGroup {subgroupname: line.sub_group_name}) \
          MATCH (group: Group{ id: line.group_id}) \
          CREATE (subgroup)-[:INCLUDED_IN]->(group) RETURN subgroup, group;").to_data_fr
```

Out[15]:

	subgroup	group
0	{'groupid': '1', 'id': '1001', 'subgroupname':...	{'id': '1', 'groupname': 'ADMIN'}
1	{'groupid': '1', 'id': '1001', 'subgroupname':...	{'id': '1', 'groupname': 'ADMIN'}
2	{'groupid': '1', 'id': '1001', 'subgroupname':...	{'id': '1', 'groupname': 'ADMIN'}
3	{'groupid': '1', 'id': '1001', 'subgroupname':...	{'id': '1', 'groupname': 'ADMIN'}
4	{'groupid': '1', 'id': '1001', 'subgroupname':...	{'id': '1', 'groupname': 'ADMIN'}
...	...	...
187	{'groupid': '22', 'id': '22003', 'subgroupname...	{'id': '22', 'groupname': 'TRANS'}
188	{'groupid': '22', 'id': '22003', 'subgroupname...	{'id': '22', 'groupname': 'TRANS'}
189	{'groupid': '22', 'id': '22003', 'subgroupname...	{'id': '22', 'groupname': 'TRANS'}
190	{'groupid': '22', 'id': '22003', 'subgroupname...	{'id': '22', 'groupname': 'TRANS'}
191	{'groupid': '22', 'id': '22003', 'subgroupname...	{'id': '22', 'groupname': 'TRANS'}

192 rows × 2 columns

In [16]:
```
# CREATE RELATIONSHIP between Users and SubGroups
gds.run("  LOAD CSV WITH HEADERS FROM 'https://raw.githubusercontent.com/letisalba/Data-
           MATCH (user: User {username: line.username}) \
           MATCH (subgroup: SubGroup{ id: line.sub_group_id}) \
           CREATE (user)-[:PART_OF]->(subgroup) RETURN user, subgroup;").to_data_frame()
```

Out[16]:

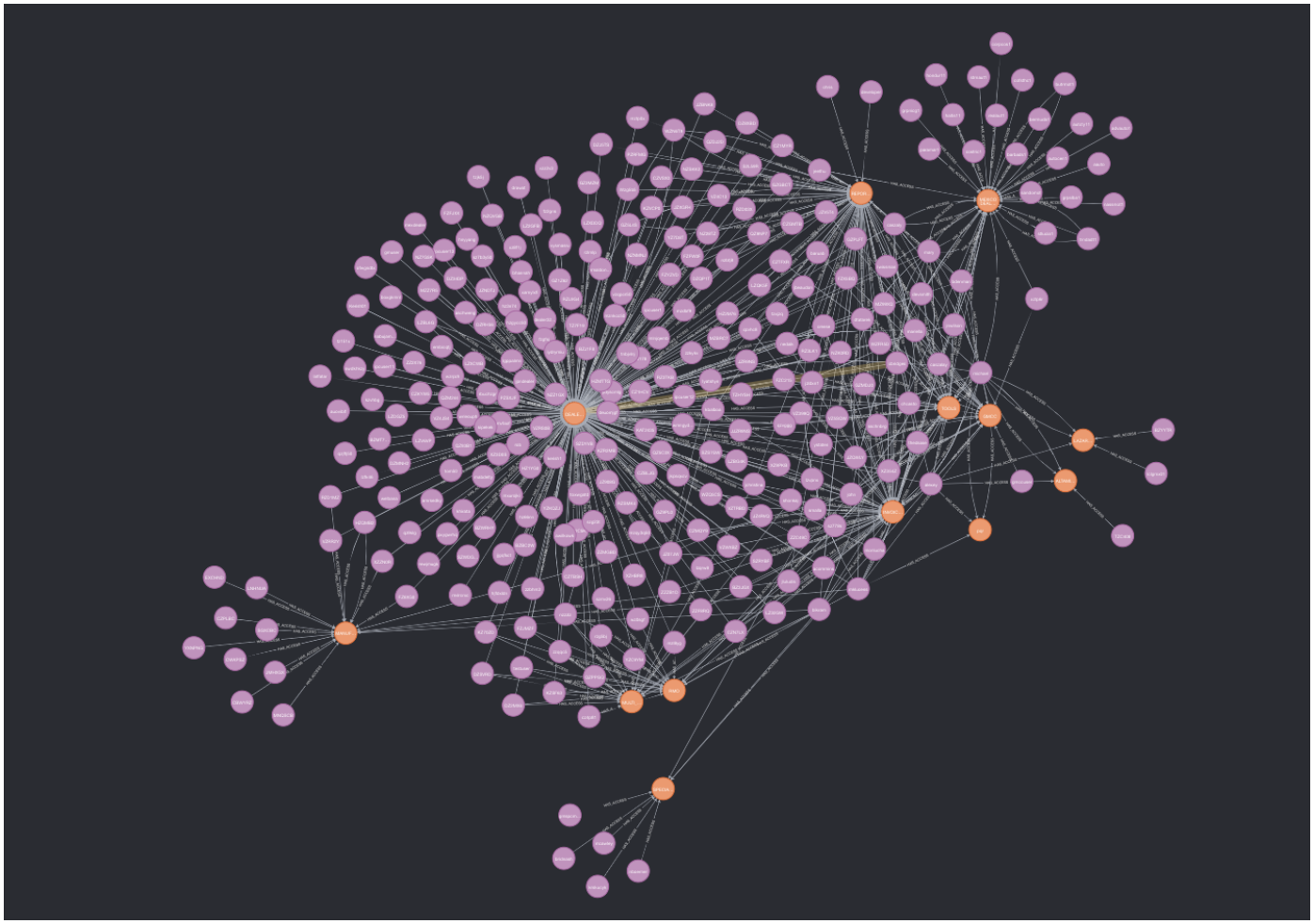
	user	subgroup
0	{'id': '233', 'username': 'bikram'}	{'groupid': '1', 'id': '1001', 'subgroupname':...
1	{'id': '233', 'username': 'bikram'}	{'groupid': '1', 'id': '1001', 'subgroupname':...
2	{'id': '233', 'username': 'bikram'}	{'groupid': '1', 'id': '1001', 'subgroupname':...
3	{'id': '233', 'username': 'bikram'}	{'groupid': '1', 'id': '1001', 'subgroupname':...
4	{'id': '233', 'username': 'bikram'}	{'groupid': '1', 'id': '1001', 'subgroupname':...
...	...	...
459	{'id': '257', 'username': 'dtatarek'}	{'groupid': '25', 'id': '25001', 'subgroupname...
460	{'id': '257', 'username': 'dtatarek'}	{'groupid': '25', 'id': '25001', 'subgroupname...
461	{'id': '257', 'username': 'dtatarek'}	{'groupid': '25', 'id': '25001', 'subgroupname...
462	{'id': '257', 'username': 'dtatarek'}	{'groupid': '25', 'id': '25001', 'subgroupname...
463	{'id': '257', 'username': 'dtatarek'}	{'groupid': '25', 'id': '25001', 'subgroupname...

464 rows × 2 columns

The first visualization is the group members connections. I've zoomed in on the graphs to see closely what these connections are comprised of.

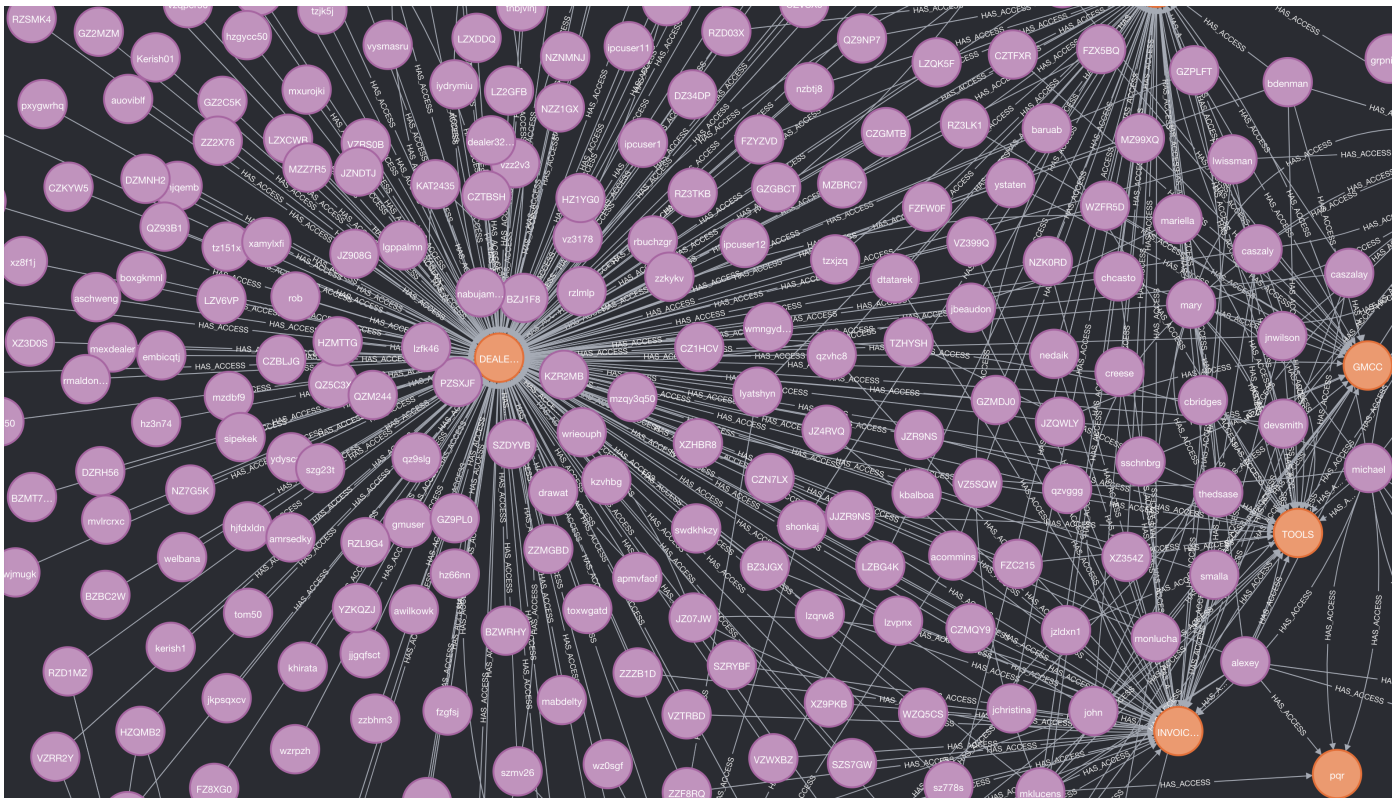
In [17]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[17]:



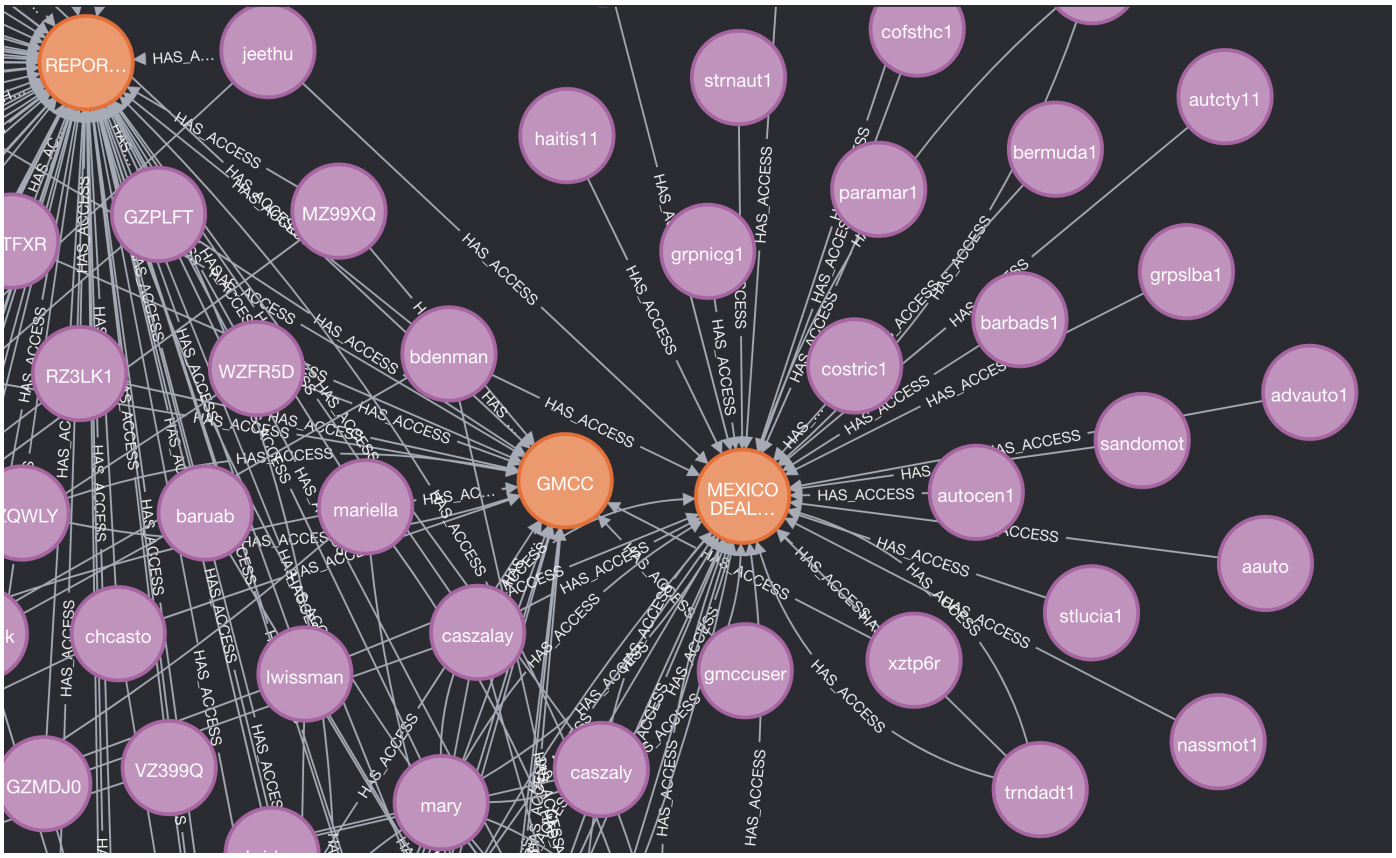
In [18]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[18]:



In [19]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

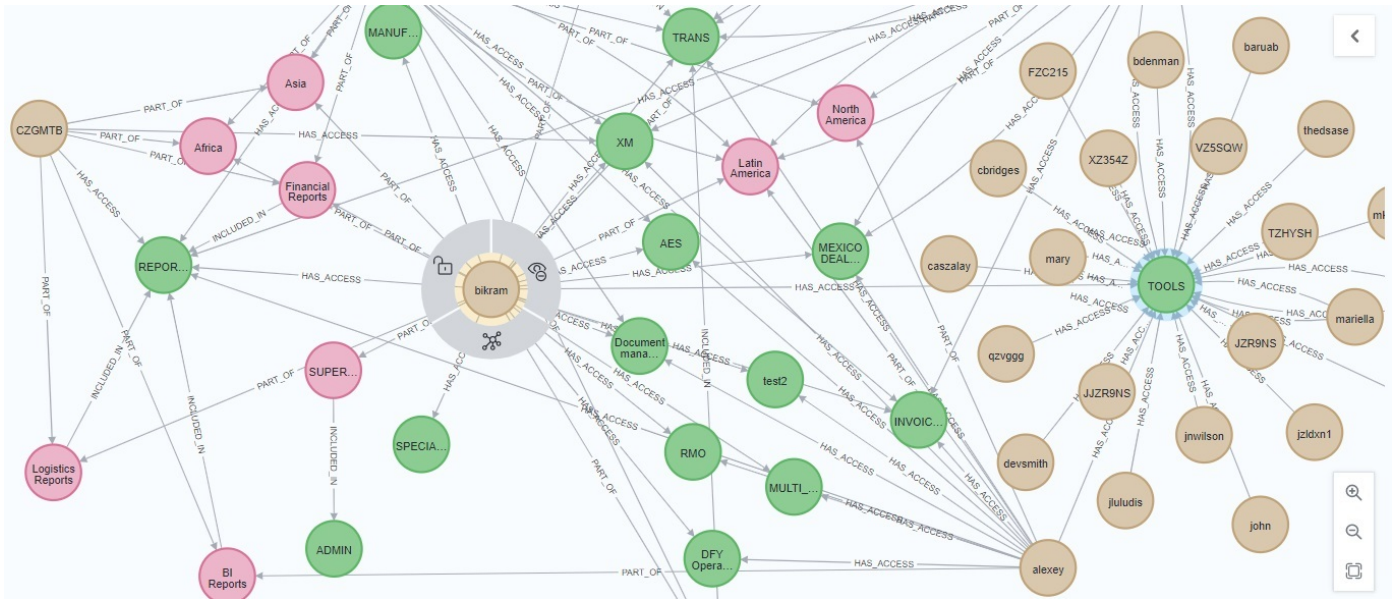
Out[19]:



Let's look at the entire relationship within the data for the subgroups

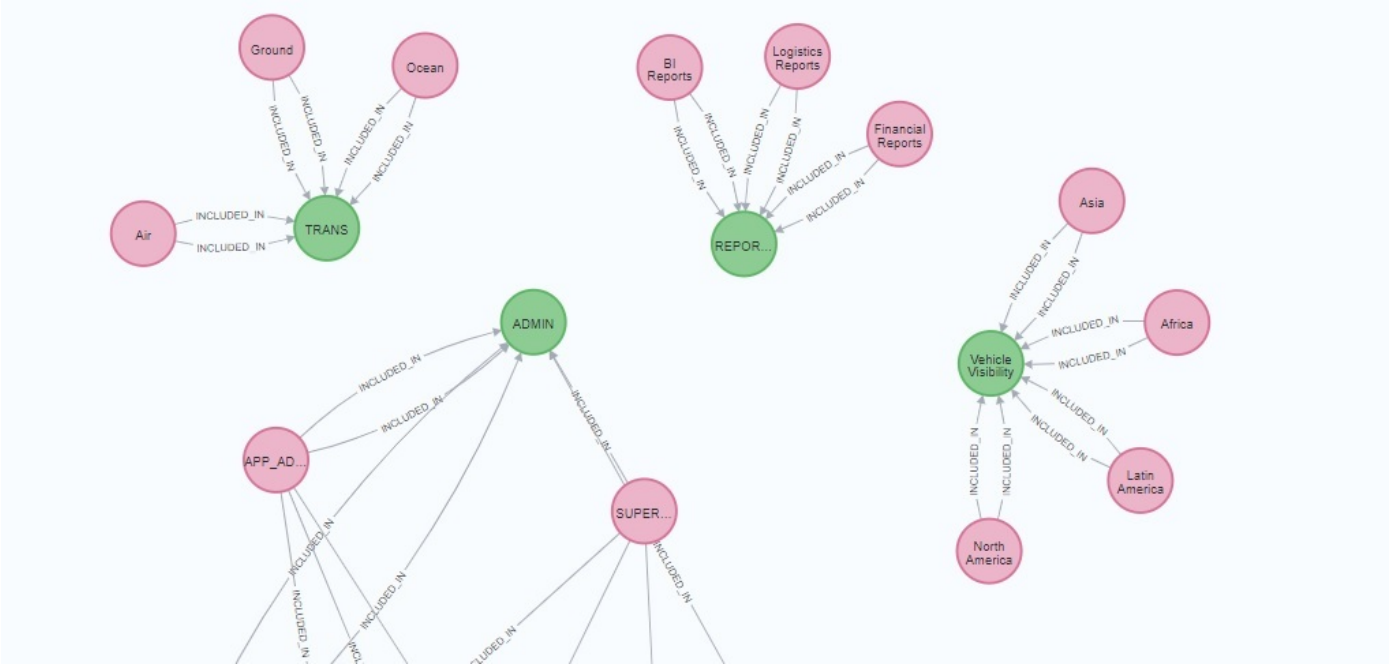
In [20]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[20]:



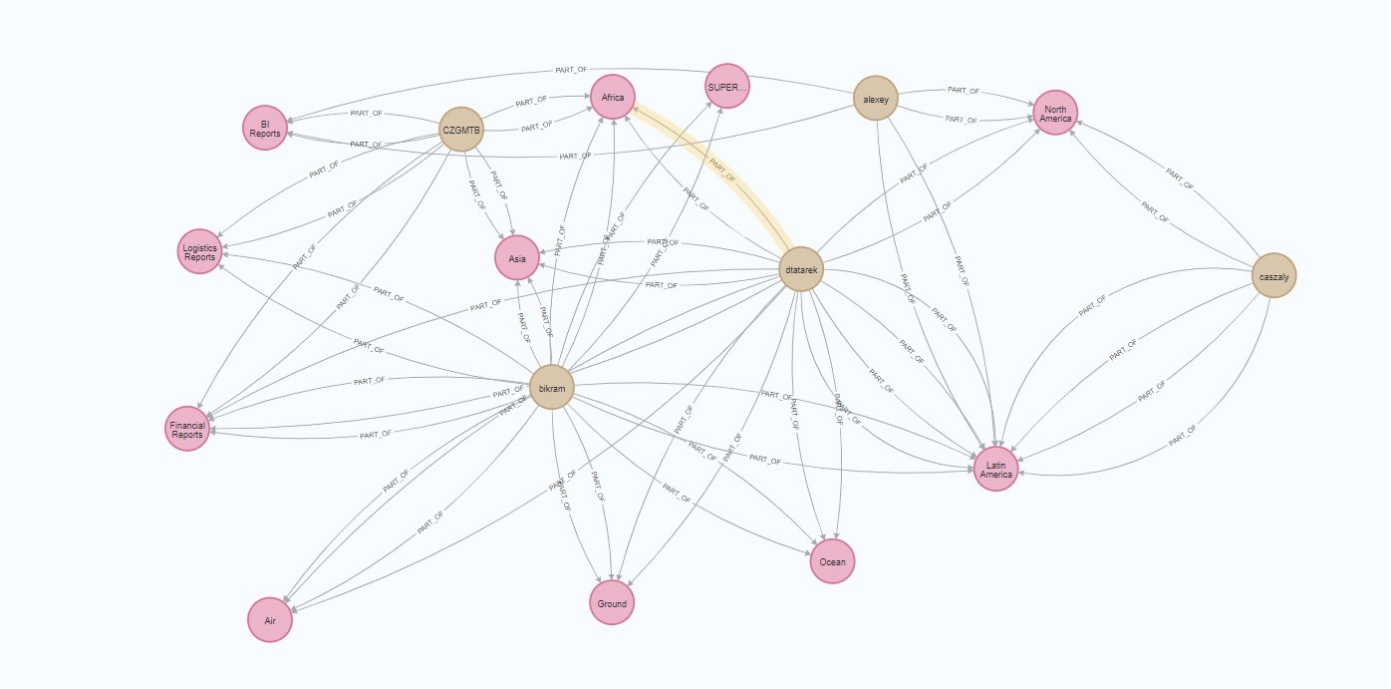
In [21]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[21]:



In [22]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

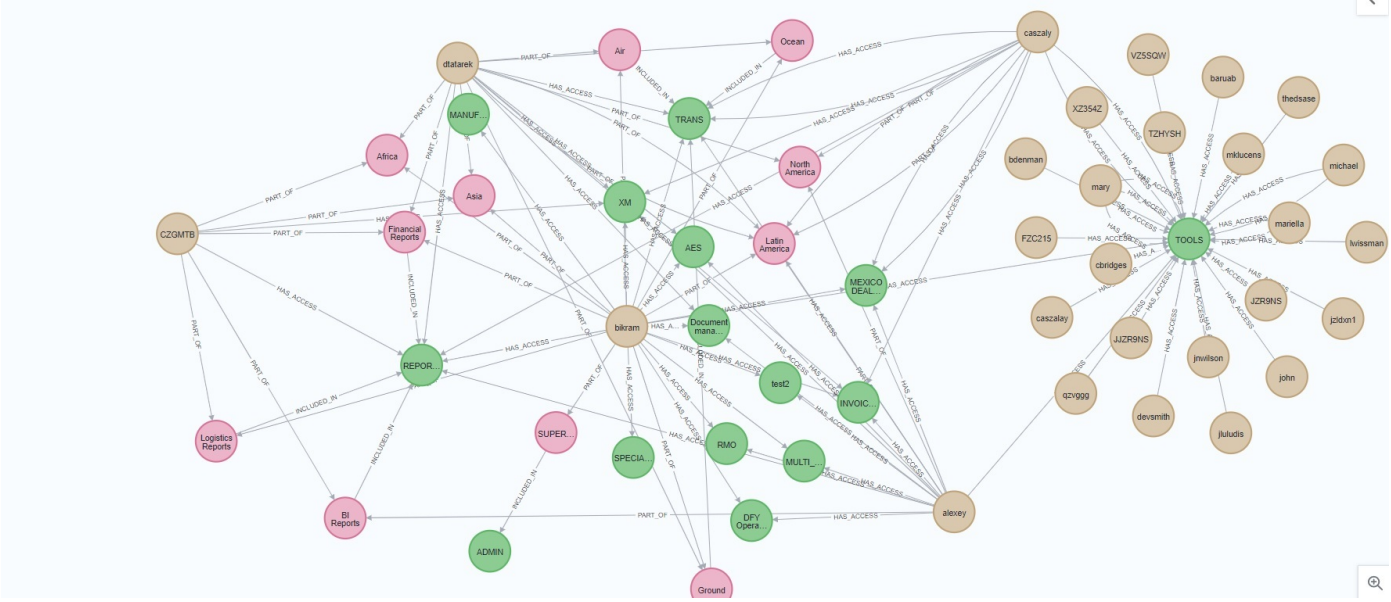
Out[22]:



These last visualizations are showing the connections of the sub-groups and user sub-groups from our data.

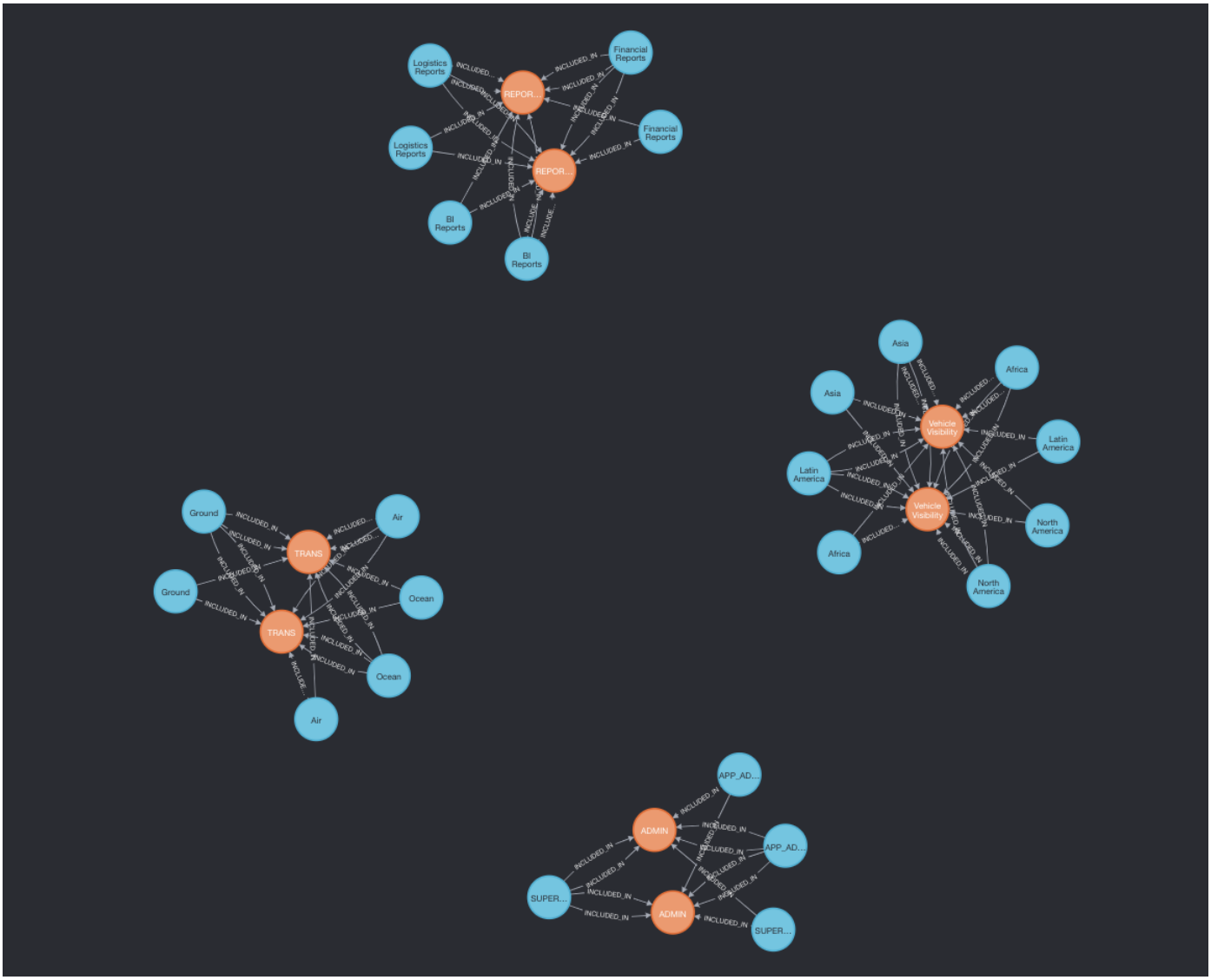
In [23]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[23]:



In [24]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[24]:

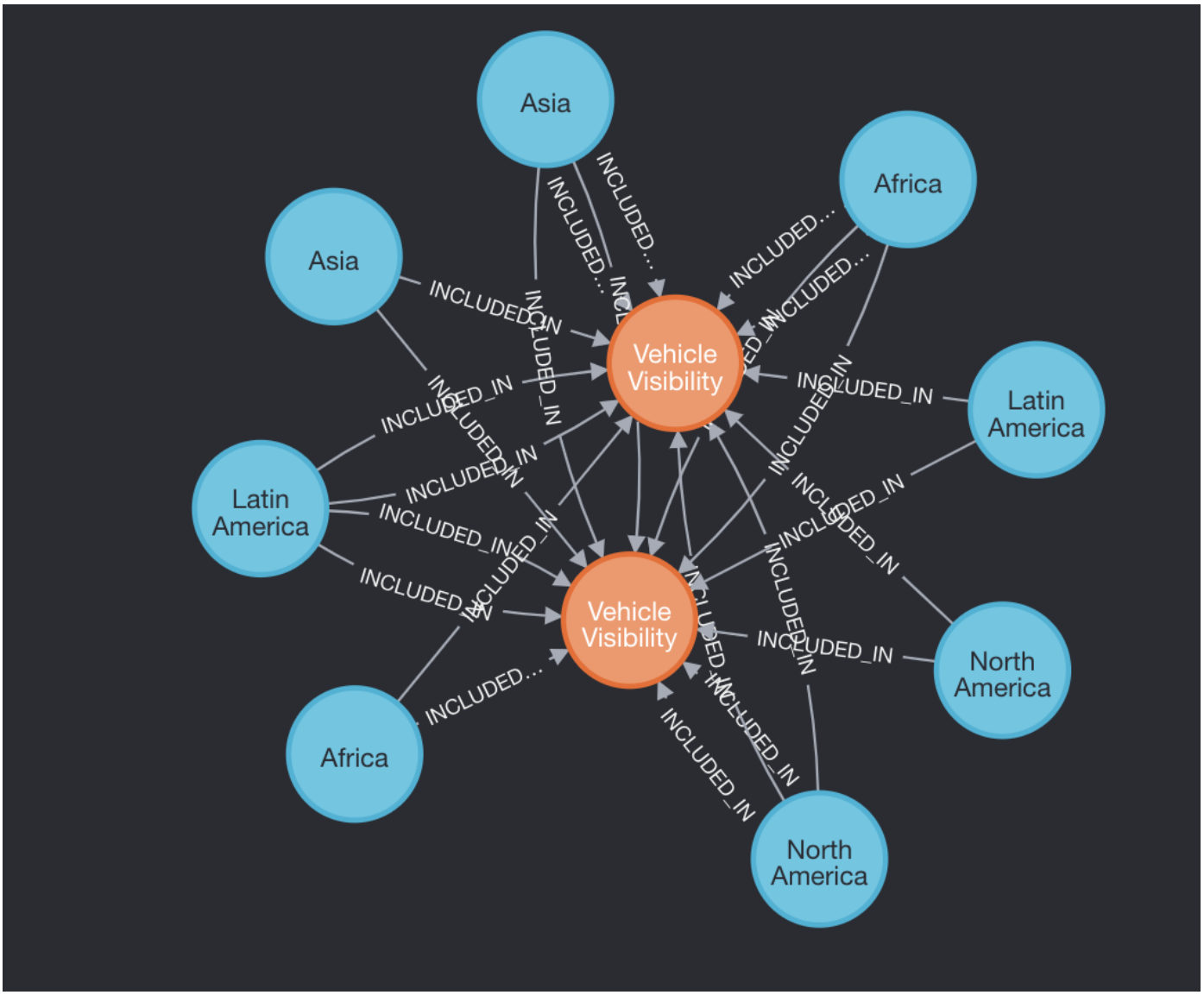


In [25]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[25]:

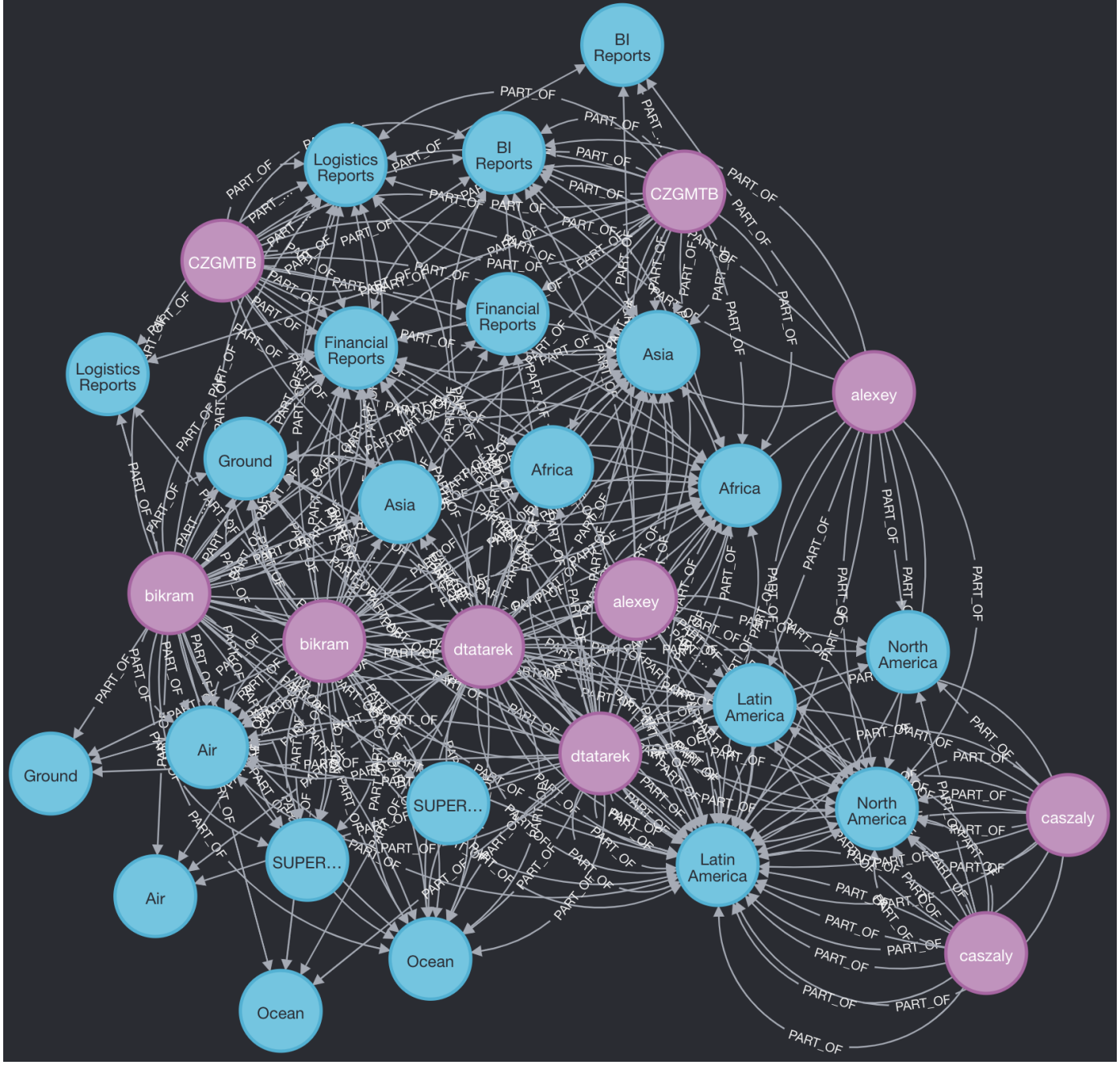
In [26]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[26]:



In [27]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[27]:



In [28]:
```
from IPython import display
display.Image("https://raw.githubusercontent.com/letisalba/Data-620/master/Week-3/pictur
```

Out[28]: