Instituto Superior Técnico

Project 1 – Complex Network Science

Grupo 27:

- André Ribeiro, nº 86384
- Diogo Ramalho, nº 86407
- Jorge Pacheco, nº 86457

Introduction

For the first project, the group decided to create and analyze a graph by choosing a dataset and converting the existing information into a graph structure, following with the calculation of some of its properties.

Dataset

The chosen dataset represents a connection of terrorist organizations with locations where the organization acted. It is a binary matrix of dimensions 394 x 65, located on the file "BAAD.csv", where the rows are indexed by the organizations and the columns are indexed by the locations. If a given entry as a 1, it means that the organization in that row acted in the location in that column. If it is 0 it means the opposite.

Extraction

Python 3 was the programming language chosen for data manipulation, mostly due to the elements of the group having previous experience with the software. The graph was structured using the NetworkX library, which provides methods for the creation, manipulation as well as import/export of graphs.

In our graph each node is an organization and nodes connect if they share a location.

With that declared, we can simplify the association process into a couple of steps:

- In each column of the array, highlight every node present on the location;
- Calculate all possible combinations between each node and associate them.

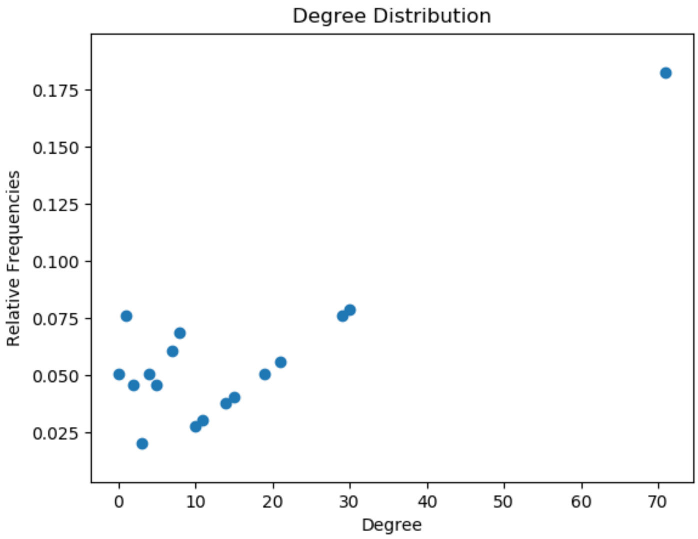
Having setup the network, the program will then export the structure into a “.gexf” file, for easier graphical analysis on the Gephi software.

Hypothesis

Before any calculations are made, one can theorize that the network will be composed by some highly clustered components (probably even islands), due to the presumption that any organization is connected to another if they share a localization. This introduces the idea that if we were to insert a node into the most active locale, it would be automatically considered as connected to all remaining components present in the same cluster. Such a premise introduces a certain margin of error, since there is a possibility of some connections being established between organizations that most likely never interacted with each other.

Analysis

In order to get a first impression of the graph, one may start by calculating its degree distribution. Such a task requires iteration over all nodes of the graph, saving its degree into a separate vector and calculate the degrees absolute frequency distribution. Before processing the data for visualization using matplotlib, each entry in the frequency vector is divided by the total amount of nodes.



We can justify the larger amount of highly connected nodes because nodes connect to each other if they share the same location, and so if a certain local is shared by many organizations, those organizations will all be connected to each other. The amount of nodes with lower degrees represent organizations that targeted less active locales. One can also notice the low variety of degree variables, given the high number of nodes present in the graph, this most likely hints at the presence of highly connected clusters described earlier, where nodes share the same degree.

- **Average Degree:** is equal to 23.07

- **SCC**: The Breadth-first search (BFS) Algorithm can be used to further prove the existence of islands. The idea is to run the algorithm over all nodes, and count the amount of different paths obtained, since different paths infer the lack of connectivity between their components. Running the algorithm returns 65 possible paths, which is the same value of the amount of different locations (columns). This happens because every organization targets exactly one location, which means that no organization will be connected to organizations in different locations.
  With this information, one can conclude that the Network is made up of 65 isolated communities, where each node is connected to each other.

- **Average path length**: the average path length is 0 for 20 locations and 1 for 45 locations. The reason for this is that in each of those 20 locations only 1 organization is active and so it will take 0 hops to go to itself. In the other 45 locations, every node is connected to every other node and so it will take 1 hop from any node to any other node.

- **Clustering coefficient**: 50 nodes have a clustering coefficient equal to 0 because these are the 50 organizations that attack in locations that have only 1 or 2 organizations active. 344 nodes have a clustering coefficient equal to 1 because they are in locations that have 3 or more organizations active and in which the organizations are all connected and in which all of your neighbors know each other. The network's clustering coefficient is equal to 0.87.

- **Centrality**:

- No node has a betweeness value, since there are no connected communities;
- Nodes in the larger communities have a higher degree centrality;
- In a given island, every node has the same closeness centrality;
- In a given island, every node has the same Katz centrality.

Bibliography

- [Source Files] https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/bigalliedanddangerousbaad